ISSN 1826-3534

federalismi.it
RIVISTA DI DIRITTO PUBBLICO ITALIANO, COMPARATO, EUROPEO

8 FEBBRAIO 2023

Censura "privata" e contrasto all'*hate speech* nell'era delle *Internet Platforms*

di Giulia Vasino
Assegnista di ricerca in Istituzioni di diritto pubblico
Sapienza – Università di Roma

Censura "privata" e contrasto all'*hate speech* nell'era delle *Internet Platforms*[*]

di Giulia Vasino
Assegnista di ricerca in Istituzioni di diritto pubblico
Sapienza – Università di Roma

Abstract [It]: Il contributo aspira ad analizzare criticamente l'attività censoria svolta dalle piattaforme digitali utilizzando il *Code of conduct on countering illegal hate speech online* come *case study*. In particolare, la riflessione intende esaminare le problematiche presenti nel Codice e il loro impatto negativo sui diritti fondamentali degli utenti. Lo scopo ultimo della disamina è valutare se tali elementi critici risultino rimodellati all'interno di un nuovo quadro normativo euro-unitario caratterizzato dalla recente adozione del DSA.

Title: "Private" censorship and countering hate speech in the age of Internet Platforms
Abstract [En]: The paper aspires to critically analyse the current issue of censorship carried out by digital platforms by using the Code of conduct on countering illegal hate speech online as a case study. In particular, the reflection intends to examine the problematic features affecting the Code and their negative impact on users' fundamental rights. The ultimate aim of the examination is to assess whether these elements have been significantly reshaped as a result of the adoption of the DSA.

Parole chiave: censura, piattaforme digitali, discorsi d'odio, Digital Services Act
Keywords: censorship, Internet platforms, hate speech, Digital Services Act

1. Introduzione: libertà di espressione e controllo del discorso pubblico nell'era delle Internet platforms

La profonda connessione e interdipendenza fra il livello di garanzia accordato alla libertà di espressione e la solidità delle radici democratiche dell'ordinamento costituisce uno dei presupposti fondanti dello

[*] Articolo sottoposto a referaggio.

Stato costituzionale[1]. La valenza "sistemica" assunta da tale diritto[2] – qualificazione cruciale dichiaratamente riconosciuta anche dal giudice delle leggi nella nota giurisprudenza che ha portato a definire la libertà di manifestazione del pensiero come «pietra angolare» del regime democratico[3] e il pluralismo informativo come «condizione preliminare» per la realizzazione e il consolidamento dei principi cardine dello stesso[4] – sembra oggigiorno conservare il suo significato prescrittivo[5], mantenendo la funzione di "banco di prova" per valutare la qualità e la "salute" della forma di Stato[6]. Se, tuttavia, i valori fondanti che sovraintendono all'esercizio di tale libertà possono definirsi immutati, le concrete modalità attraverso le quali essa si manifesta mutano progressivamente forma in una realtà caratterizzata dalla pervasività di Internet e dei *social network*[7]. È innegabile, infatti, come i canali tradizionali appaiano, nell'odierno contesto, profondamente ridefiniti[8], e siano i cosiddetti "Stati dai confini immateriali"[9], rappresentati dalle grandi piattaforme digitali, a costituire veicoli privilegiati del pensiero individuale e

[1] Non perde la sua attualità e centralità la nota riflessione di Carlo Esposito concernente il nesso fra libertà di manifestazione del pensiero, quale diritto individuale, e la vita democratica dell'ordinamento, in C. ESPOSITO, *La libertà di manifestazione del pensiero nell'ordinamento italiano*, Milano, Giuffrè, 1958, pp. 11-12. Cfr., altresì, la "speculare" e coeva argomentazione di Emerson sulla centralità della libertà di espressione all'interno dei sistemi democratici: «(…)*the right of all members of society to form their own beliefs and communicate them freely to others must be regarded as an essential principle of a democratically-organized society (…). This is, of course, especially true of political decisions. But the basic theory carried beyond the political realm. It embraced the right to participate in the building of the whole culture, and included freedom of expression in religion, literature, art, science and all areas of human learning and knowledge*», in T. EMERSON, *Toward a General Theory of the First Amendment*, in *Yale Law Journal*, vol. 72, 1963, p. 883.

[2] Cfr. l'autorevole ricostruzione di P. BARILE, *Diritti dell'uomo e libertà fondamentali*, Bologna, Il Mulino, 1984, il quale si sofferma proprio sulla duplice dimensione del diritto alla libera manifestazione del pensiero, dalla quale emerge una vocazione "funzionale" di tale diritto fondamentale (pp. 228-229); v., inoltre, L. PALADIN, *Libertà di pensiero e libertà di informazione: le problematiche attuali*, in *Quad. cost.*, n. 1, 1987, p. 5 ss.

[3] Così la Corte costituzionale nella sentenza n. 84 del 1969, espressione seguita successivamente dall' esplicativa definizione della libertà di manifestazione del pensiero come «cardine di democrazia dell'ordinamento generale» (sent. n. 126 del 1985). Rimane altrettanto potente l'inquadramento della libertà di espressione come diritto «coessenziale al regime di libertà garantito dalla Costituzione» (sent. n. 11 del 1968).

[4] Sul collegamento fra pluralismo informativo e attuazione dei principi democratici, oltre alla nota sentenza n. 29 del 1996, la Corte ritorna sul punto, *ex multis*, nelle decisioni nn. 234 del 1990, 21 del 1991, 826 del 1988, 112 del 1993, 155 del 2002, 312 del 2003. Nella giurisprudenza più recente, inoltre, la Consulta ha ribadito la permanente attualità di tali legami, ribadendo come «il "diritto all'informazione", poi, secondo l'insegnamento di questa Corte, va determinato e qualificato in riferimento ai princìpi fondanti della forma di Stato delineata dalla Costituzione, i quali esigono che la nostra democrazia sia basata su una libera opinione pubblica e sia in grado di svilupparsi attraverso la pari concorrenza di tutti alla formazione della volontà generale» (sent. 206 del 2019, punto 5.2 del *Considerato in diritto*).

[5] Inoltre, tale peculiare qualificazione è stata sugellata anche dalla Corte di Strasburgo, la quale ha definito la libertà di espressione come «una delle essenziali fondamenta di una società democratica» nonché «una delle condizioni basilari per lo sviluppo della persona umana», cfr. Corte Edu, *Handyside c. Regno Unito*, 1976.

[6] Per una compiuta ricostruzione di tale diritto, sotto il profilo dottrinale e giurisprudenziale, e in un'ottica comparatistica, si veda l'autorevole studio curato da M. LUCIANI, *La libertà di espressione, una prospettiva di diritto comparato – Italia*, EPRS-Servizio Ricerca del Parlamento europeo, ottobre 2019, p. 9 ss. e 33 ss.

[7] Cfr. Y. BENKLER, *The Wealth of Networks: How social production transforms markets and freedom*, Yale University Press, 2006.

[8] V. L. CALIFANO, *La libertà di manifestazione del pensiero ... in rete; nuove frontiere di esercizio di un diritto antico. Fake news, hate speech e profili di responsabilità dei social network,* in *federalismi.it*, fasc. 26/2021, pp. 3-4.

[9] G.L. CONTI, *Manifestazione del pensiero attraverso la rete e trasformazione della libertà di espressione: c'è ancora da ballare per strada?*, in *Rivista AIC*, 4/2018, p. 202.

collettivo[10]. Di conseguenza, le categorie teoriche classiche che circoscrivono tale diritto essenziale devono oramai fare i conti con nuovi "attori protagonisti" della libertà di espressione su scala globale, soggetti in grado di modellare, gestire e influenzare profondamente il discorso pubblico[11].

Di fronte a una realtà in evoluzione, appare forse superfluo sottolineare, da un lato, come la disponibilità di tali strumenti tecnologici accanto ai mezzi tradizionali abbia consentito, al contempo, di massimizzare l'accesso all'informazione, nonché di diffondere contenuti, idee, opinioni con una rapidità sconosciuta e mediante formule sconfinate[12].

D'altro canto, tuttavia, è ormai da tempo in corso un processo di presa di coscienza dei peculiari rischi connessi a tale modalità solo "apparentemente disintermediata" di esercizio della libertà di espressione in presenza di "colossi digitali" assimilabili sempre di più a "*new governors*" dell'informazione[13]. In tale scenario, dunque, pur mantenendo fermo l'indubitabile impatto positivo che la rivoluzione tecnologica ha determinato in termini di democratizzazione e universalizzazione di tale diritto[14], si prende atto del fatto che quelle utopiche velleità ultra-libertarie e anarchiche[15] proiettate verso la realizzazione di una realtà virtuale in cui potesse effettivamente affermarsi un "*truly free speech*"[16] abbiano da tempo perduto la loro capacità persuasiva[17]. Difatti, l'attività di moderazione – intensa, *lato sensu*, come quell'azione di organizzazione, gestione, vaglio e rimozione dei contenuti ospitati dal *provider* – svolta in ottemperanza ad una cornice regolatoria implementata dal *social network*, rappresenterebbe oggi l'attività che connota

[10] Di particolare impatto risulta la metafora del "triangolo" coniata da Balkin per indicare il passaggio, nella sfera della libertà di espressione, da un modello bilaterale basato sulla relazione Stato-cittadino a un modello trilaterale in cui si affaccia un nuovo soggetto, la piattaforma digitale, in J.M. BALKIN, *Free Speech is a Triangle*, in *Columbia Law Review,* vol. 118, n. 7, 2018, p. 2011 ss.

[11] Cfr. K. KLONICK, *The new governors: the people, rules, and processes governing online speech*, in *Harvard Law Review*, vol. 131, n.6, 2018, p. 1662 ss.

[12] V. E. VOLOKH, *Cheap Speech and What It Will Do*, in *The Yale Law Journal* , vol. 104, no. 7, 1995, p. 1833.

[13] J. BALKIN, *Old-school/ new-school speech regulation*, in *Harvard Law Review*, vol. 127, 2014, argomenta, infatti: «*a widely noted and characteristic feature of the digital age is the democratization of information production, and therefore the democratization of opportunities to speak and express one's self. The "disintermediation" often associated with the Internet does not involve the abolition of media gatekeepers but rather the substitution of one kind of infrastructure for another*» (p. 2304).

[14] A. MICHAEL FROOMKIN, *Habermas@Discourse.net: Toward a Critical Theory of Cyberspace*, in *Harvard Law Review*, vol. 116, n. 3, 2003, pp. 856-857.

[15] Cfr. O. POLLICINO- M. BASSINI, *The Law of the Internet Between Globalisation and Localisation*, in M. MADURO et AL. (a cura di), *Transnational Law: Rethinking European Law and Legal Thinking*, Cambridge, Cambridge University Press, 2014, p. 346 ss.

[16] In altre parole, dunque, nell'era della complessità digitale, l'auspicio positivo per il quale Internet sarebbe divenuto il «facilitatore del "libero mercato delle idee», espresso nella ormai celebre sentenza *Reno v. ACLU* (1997) della Corte suprema americana apparirebbe oggi un'espressione che richiede di essere utilizzata con estrema cautela. Sul punto cfr. M. MANETTI, *Regolare l'Internet*, in *Medialaws.eu*, 2/2020, p. 36.

[17] Pur non potendo dedicare spazio, in questa sede, ai profili relativi all'inquadramento normativo di Internet e, più specificamente, dei *social network*, si rimanda alla ricostruzione di M. BASSINI, *Libertà di espressione e social network, tra nuovi "spazi pubblici" e "poteri privati". Spunti di comparazione*, in *Rivista italiana di informatica e diritto*, 2/2021, p. 45, il quale argomenta come, per lungo tempo, sia in Europa sia negli Stati Uniti, sia prevalso un approccio regolatorio «volutamente minimalista», cioè volto a non irregimentare il funzionamento delle piattaforme all'interno di un sistema normativo eccessivamente rigido, proprio alla luce della condivisa concezione che tale prospettiva avrebbe avuto potenzialità espansive per la circolazione di idee e contenuti.

"ontologicamente" le piattaforme digitali[18]. Dunque, nonostante l'auto-qualificazione neutralista che le *Internet platforms* stesse tendono sovente ad attribuirsi[19], e malgrado la definizione di *hosting* che rimane nella disciplina continentale attualmente vigente, i grandi attori del cyberspazio rappresentano tutt'altro che meri contenitori dell'informazione ma, al contrario, si connotano oramai come potenti soggetti attivi in grado di decidere della permanenza o eliminazione di un determinato contenuto sulla propria piattaforma[20]. In altre parole, dunque, i custodi della sfera virtuale agirebbero oggi alla stregua di veri e propri censori privati, la cui azione è in grado di affiancarsi a quella del decisore pubblico, con indubbie conseguenze di rilievo costituzionale[21].

Infatti, se, proprio in forza della summenzionata co-essenzialità fra libertà di espressione e forma di Stato, la censura costituisce già di per sé un fenomeno "spinoso" agli occhi dell'ordinamento[22], a maggior

[18] Sottolinea questa peculiare configurazione T. GILLESPIE, *Custodians of the Internet: platforms, content moderation, and the hidden decisions that shape social media*, New Haven-London, Yale University Press, 2018, p. 5; l'A. argomenta come le *Internet Platforms*, dopo aver "conquistato" lo spazio online, si siano trasformate in custodi dello stesso, rivestendo contemporaneamente il ruolo di legislatori, interpreti e giudici della dimensione virtuale.

[19] Si pensi, ad esempio, all'assenza di *policies* particolarmente limitanti la libertà di espressione voluta da Twitter. Tale piattaforma, infatti, ha inteso perseguire una rigida linea di neutralità per lungo tempo, autodefinendosi «*the free speech wing of the free speech party*», cfr. J. HALLIDAY, *Twitter's Tony Wang: "We Are the Free Speech Wing of the Free Speech Party"*, in *The Guardian*, 22 marzo 2012.

[20] Risultano esemplificative dell'importanza di tale ruolo, anche a causa delle diverse interpretazioni emerse in dottrina sulle questioni oggetto di sindacato, proprio le recenti vicende giurisdizionali interne, sviluppatesi a partire dal 2019, relative all'oscuramento degli account di Forza Nuova e Casa Pound da parte di Facebook, provvedimenti che hanno condotto all'adozione di pronunce differenti in sede cautelare da parte del Tribunale di Roma. In ambedue le pronunce emerge una delineazione differente della qualificazione del soggetto privato e della funzione che esso può assumere in presenza di contenuti odiosi, potenzialmente suscettibili di ledere i valori costituzionali, provenienti da un soggetto politico. Nella vicenda concernente Casa Pound (Trib. Roma, ord. 12 dicembre 2019) l'obbligo di ripristino della pagina viene ricondotto al ruolo peculiare svolto dalla piattaforma privata e alla sua natura di "foro pubblico" che imporrebbe ad essa di operare con responsabilità e in ottemperanza ai principi repubblicani: la rimozione di un partito politico dalla piattaforma costituirebbe un attentato al principio pluralistico. La valutazione della eventuale contrarietà dei fini dell'Associazione con i valori fondanti del sistema andrebbe accertata in altra sede. Diversamente, in occasione di un'analoga vicenda concernente Forza nuova (Trib. Roma, ord. 23 febbraio 2019), l'inquadramento in senso pubblicistico della funzione svolta dal soggetto privato conduce l'autorità giurisdizionale a legittimare il blocco dell'account in quanto il gestore sarebbe vincolato a dar seguito alle prescrizioni costituzionali e sovranazionali concernenti il contrasto ai discorsi d'odio. Quest'ultima posizione è stata ripresa di recente nella decisione di merito (sentenza di primo grado n. 17909 del 5 dicembre 2022, Tribunale di Roma) con la quale si è chiusa la controversia fra Casapound e Facebook mediante la revoca dell'ordinanza cautelare e la conferma della liceità della condotta della piattaforma la quale, nella visione del collegio, avrebbe agito in applicazione delle proprie regole contrattuali, sottoscritte dall'utente, avendo altresì il dovere di rimuovere contenuti illeciti secondo la normativa europea e internazionale. Per una differente lettura delle prime pronunce - soprattutto con specifico riferimento alla diversa cogenza che assumerebbero le norme della CEDU in tale ambito a carico del soggetto pubblico e del soggetto privato – si rimanda a C. CARUSO, *I custodi di silicio. Protezione della democrazia e libertà di espressione nell'era dei social network*, in *Consultaonline.org*, 17 marzo 2020, p. 1 ss. il quale parla di «giudizi ad alta tensione assiologica» e, altresì, a P. DE SENA E M. CASTELLANETA, *La libertà di espressione e le norme internazionali, ed europee, prese sul serio: sempre su Casapound c. Facebook,* in *Sidiblog.org*, 20 gennaio 2020.

[21] M. MANETTI, *Facebook, Trump e la fedeltà alla Costituzione*, in *Forum di Quaderni costituzionali*, n. 1, 2021.

[22] Com'è noto, si tratta di un solido baluardo garantistico che accomuna la tradizione giuridica occidentale ma che presenta tuttavia delle importanti differenze sostanziali se si compara l'approccio continentale a quello statunitense. In quest'ultima, infatti, la centralità assiologica del primo emendamento imporrebbe che ogni forma di limitazione della libertà d'espressione prevista per via normativa sia sottoposta a uno *strict scrutiny* e da tale test chiave difficilmente risulterebbe legittima. Sul punto cfr. A. BALDASSARRE, *Privacy e Costituzione. L'esperienza statunitense*, Roma, Bulzoni, 1974,

ragione risulterà incompatibile con i principi fondanti dello Stato costituzionale l'affidamento di tale delicata funzione a soggetti privati la cui primigenia natura rimane quella di operatori economici[23]. In un'epoca in cui, quindi, ancora si discute delle soluzioni più corrette ed efficaci per inquadrare il cyberspazio in una cornice di legalità costituzionale[24], l'attività di moderazione esercitata dalle *Internet platforms* rappresenta oggigiorno uno dei risvolti più tangibili e preoccupanti della erosione delle funzioni tradizionali dello Stato di fronte al ruolo sempre più pervasivo e "totalizzante" assunto dai nuovi poteri

p. 192 ss. Alle spalle di tale approccio si staglia una lettura "assolutista" della libertà di espressione, affermatasi soprattutto fra il 1940 e il 1970, che, come autorevolmente sottolineato, si fonda su 4 fondamentali idee chiave: «*the first idea is that the government is the enemy of freedom of speech. Any effort to regulate speech, by the nation or the states, is threatening to the principle of free expression (…). The second idea is that we should understand the First Amendment as embodying a commitment to a certain form of neutrality. Government may not draw lines between speech it likes and speech it hates (…). The third idea is that we should not limit the principle of free expression to political speech, or to expression with a self-conscious political component. The final idea is that any restrictions on speech, once permitted, have a sinister and inevitable tendency to expand. (…)"Slippery slope" arguments therefore deserve a prominent place in the theory of free expression*», in C. R. SUNSTEIN, *Free Speech*, in *The University of Chicago Law Review*, vol. 59, n. 1, "*The Bill of Rights in the Welfare State: A Bicentennial Symposium* (Winter)", 1992, pp. 259-260. Tale impostazione garantistica si coglie anche se si rivolge l'attenzione alla formulazione dell'articolo 21 secondo comma della nostra Costituzione, il quale esclude esplicitamente che la stampa possa essere soggetta ad autorizzazione e censura. Difatti, come chiarito dalla nostra Consulta «il divieto di cui all'art. 21, secondo comma, della Costituzione concerne la censura quale istituto tipico del diritto pubblico, secondo cui gli organi dello Stato, e soltanto essi, esercitano autoritativamente un controllo preventivo» (Corte cost., sent. 93 del 1972, *Considerato in diritto*). La disposizione costituzionale prevede soltanto, al terzo comma, l'ipotesi del sequestro successivo alla pubblicazione, il quale può aver luogo in determinate circostanze e in presenza delle note garanzie della riserva di legge e di giurisdizione. Sul punto si rimanda alla giurisprudenza chiarificatrice della Corte costituzionale in materia, v., *ex multis*, sentt. 115 del 1957, 44 del 1960, 4 del 1972, 92 del 1979.

[23] Cfr. la riflessione di M. CUNIBERTI, *Potere e libertà nella rete*, in *Medialaws.eu*, 3/2018, p. 39 ss. il quale, concentrandosi sulla teoria, emersa in dottrina, sull'assimilazione del *social network* ad una "formazione sociale", invita a mantenere ben presenti le differenze radicali fra le due entità, divergenze motivate in primo luogo dalla alterità di scopo riguardanti i membri della *community*, da un lato, e il gestore, dall'altro, il quale mantiene una configurazione proprietaria e privatistica. Si veda l'inquadramento di tale problematica, sotto il profilo costituzionalistico, elaborata da M. MONTI, *Le Internet platforms, il discorso pubblico e la democrazia*, in *Quaderni costituzionali*, 4/2019, p. 811 ss.; per una disamina più ampia concernente il tema posto sullo sfondo, ovvero l'interazione, l'intreccio e la commistione fra diritto pubblico e privato e, più specificamente, il tema del "*constitutionalization of private law*", si rimanda a O. GERSTENBERG, *Private Law and the New European Constitutional Settlement,* in *European Law Journal,* vol. 10, n. 6, November 2004, pp. 766 ss.

[24] Si allude, in senso lato, sia al fenomeno del cosiddetto "costituzionalismo digitale", inteso come l'insieme di iniziative volte a creare un solido *framework* normativo che ponga dei limiti all'esercizio del potere in Internet nell'ottica di rendere il funzionamento della dimensione virtuale compatibile con i principi dello Stato di diritto e con le istanze di tutela dei diritti fondamentali. Tale definizione è stata resa nota dalla celebre ricerca pubblicata dall'università di Harvard, si veda L. GILL, D. REDEKER, U. GASSER, *Towards Digital Constitutionalism? Mapping Attempts to Craft an Internet Bill of Rights*, in *Berkman Center Research Publication*, n. 15, 2015. Per una ricostruzione del percorso politico e normativo che ha condotto l'Unione europea verso questo approccio si rimanda alla ricostruzione di G. DE GREGORIO, *The rise of digital constitutionalism in the European Union*, in *International Journal of Constitutional Law*, vol. 19, issue 1, p. 41 ss. il quale, nel ripercorrere la strada che ha condotto al mutamento di paradigma da parte del legislatore continentale, da una prospettiva economicistica a una "*rights-oriented*" individua una pluralità di fattori determinanti: uno fattuale, quale il progressivo ruolo economico assunto dalle piattaforme digitali, uno giuridico, quale la "costituzionalizzazione" della Carta di Nizza, con la conseguente elevazione e cogenza dei principi in essa contenuti, nonché un fattore istituzionale, quale il ruolo chiave della Corte di Giustizia dell'Unione nel traghettare il sistema verso un nuovo assetto proiettato verso la tutela dei diritti fondamentali. Per un approfondimento in merito alla pluralità di interventi e ai criteri ispiratori che possono essere ricondotti a tale filone cfr. M. SANTANIELLO, *Sovranità digitale e diritti fondamentali: un modello europeo di Internet governance*, in *Rivista italiana di informatica e diritto*, 1/2022, pp. 47-48.

del *Web*. Tale tensione emerge, soprattutto, nelle circostanze in cui il ruolo assunto dagli stessi diventi non meramente ancillare ma surrogatorio rispetto a quello esercitato dal decisore pubblico[25].

2. Il *Code of conduct on countering illegal hate speech online* come modello problematico di "censura privata"

In merito alla moderazione effettuata dalle piattaforme digitali è opportuno, infatti, tracciare adeguate distinzioni. Invero, quando si allude alla progressiva traslazione dell'esercizio del potere di censura nelle mani delle nuove autorità private[26], si allude a diverse modalità di intervento. Da un lato, si fa riferimento a quella attività svolta in forza di un quadro normativo stabilito dal *social* stesso, i cosiddetti "standard" o "linee guida" che l'utente sottoscrive al momento dell'iscrizione. Quest'ultime costituirebbero un insieme di regole autonomamente definite dal singolo *provider* e istituite con il precipuo scopo di rendere lo spazio virtuale un *safe environment* per i propri membri, essendo finalizzate principalmente a prevenire la diffusione di contenuti violenti o discriminatori. Si tratterebbe, dunque, di un *framework* garantistico non uniforme né sotto il profilo delle restrizioni applicate alla libertà di manifestazione del pensiero né sotto il profilo degli strumenti utilizzati[27], e che lascia spazio ad ampi margini di arbitrarietà per la piattaforma nella fase esecutiva[28]. Tuttavia, tale sub-categoria farebbe insorgere limitate problematiche sotto il profilo costituzionalistico, trattandosi di un'attività di *content moderation* che le *Internet* platforms realizzano in assenza di indicazioni provenienti dall'autorità statuale. Il soggetto, infatti, darebbe esclusiva applicazione alle summenzionate regole vigenti all'interno della propria *community*, qualificabile alla stregua di un ordinamento privatistico. All'interno di quest'ultimo, dunque, il gestore attuerebbe una censura "interna", cioè svolta in assenza di una previsione legislativa statuale e/o una decisione dell'autorità giudiziaria alla base che qualifichi il contenuto come illecito.

Ciò che tuttavia acquista maggior rilevanza in un'analisi volta ad indagare il rimodellamento del rapporto fra soggetto pubblico e privato, è la cosiddetta censura *de jure*, identificabile come quell'attività svolta *in tandem* fra Stato e potere privato ai fini della rimozione di contenuti illeciti[29]. In tale circostanza, dunque, *l'Internet platform* andrebbe a rivestire i panni di *longa manus* dell'autorità statuale in presenza di una

[25] S. F. KREIMER, *Censorship by Proxy: The First Amendment, Internet Intermediaries, and the Problem of the Weakest Link*, in *University of Pennsylvania Law Review*, 2006, in particolare pp. 13-14 e 28-27.

[26] Cfr. G. RESTA, *Diritti fondamentali e diritto privato nel contesto digitale,* in G. RESTA-F. CAGGIA (a cura di), *I diritti fondamentali in Europa e il diritto privato,* Roma, RomaTre Press, 2019, p. 117 ss.

[27] Questi meccanismi di moderazione, in particolare quelli di *hard moderation*, si basano su sistemi di rilevazione di tipo automatizzato (AI) o umano, solitamente ibrido, su meccanismi di natura preventiva o, più comunemente, successiva e sulle segnalazioni degli utenti. Per un approfondimento circa le differenti modalità attraverso le quali la piattaforma modera v. J. GRIMMELMANN, *The Virtues of Moderation*, in *Yale Journal of Law and Technology*, n. 17, 2015, p. 61 ss.

[28] Cfr. J. ROSEN, *Who Decides? Civility v. Hate Speech on the Internet*, in *Insights on Law & Society*, n. 2, 2013.

[29] Si riprende la classificazione e la terminologia utilizzata da M. MONTI, *Privatizzazione della censura e Internet platforms: la libertà d'espressione e i nuovi censori dell'agorà digitale*, in *Rivista italiana di informatica e diritto*, 1/2019, p. 35 ss.

violazione, ponendo a fondamento legale della propria azione una disciplina legislativa e/o una decisione riconducibile ad un organo amministrativo o giurisdizionale. In altre parole, quindi, il decisore pubblico si servirebbe della cooperazione della piattaforma per ripristinare la legalità nella dimensione digitale[30].

È opportuno premettere, però, che dai modelli normativi in cui il *provider* agisce come mero esecutore di una decisione rimessa interamente al decisore pubblico scaturiscano limitate criticità[31]. Diversamente, invece, maggiori tensioni affiorano nel momento in cui è lo Stato ad affidare alla piattaforma un vero e proprio compito di "censura sostanziale", delegando a quest'ultima valutazioni discrezionali in merito alla liceità dei contenuti[32]. Appare evidente come, in questa ipotesi, un soggetto privato, la cui primigenia natura rimane quella di operatore economico, diventi titolare di un considerevole potere decisorio nonché del compito di operare delicati bilanciamenti fra interessi contrapposti[33].

In tale direzione, caratterizzata dal progressivo protagonismo del *social media* nell'inquadramento della fattispecie, pare in realtà procedere sia il legislatore euro-unitario sia nazionale[34]. Se si svolge lo sguardo alla dimensione continentale, segue tale schema di contrasto dei discorsi d'odio il *Code of conduct on countering illegal hate speech online* (Codice di condotta per lottare contro la diffusione dell'incitamento all'odio online). Pur trattandosi di un atto a carattere non vincolante, come si evince dalla sua denominazione, tale disciplina affida alle piattaforme il compito di eliminare contenuti qualificabili come atti di

[30] Cfr. J. BALKIN, *Old-school/new-school speech regulation*, cit., il quale identifica in tale sinergia pubblico-privato il tratto costitutivo del nuovo metodo di regolazione del discorso pubblico: «*public/private cooperation and co-optation are hallmarks of new-school speech regulation. To the extent that the government does not own the infrastructure of free expression, it needs to coerce or co-opt private owners to assist in speech regulation and surveillance - to help the state identify speakers and sites that the government seeks to watch, regulate, or shut down. To this end, the government may offer a combination of carrots and sticks, including legal immunity for assisting the government's efforts at surveillance and control*» (p. 2299).

[31] Si parla, in tale ipotesi, di una censura meramente "funzionale", v. O. GRANDINETTI, *Le piattaforme digitali come "poteri privati" e la censura online*, in *Rivista italiana di informatica e diritto*, 1/2022, p. 181. Volendo volgere lo sguardo all'interno del nostro ordinamento, segue tale meccanismo il contrasto alla pedopornografia online. Sulla base dell'art. 14*bis* l. n. 269/1998, il Centro nazionale per il contrasto della pedopornografia fornisce una lista di siti pedopornografici agli ISP affinché esse provvedano al blocco delle pagine. La blacklist è soggetta ad un aggiornamento periodico svolto sotto il controllo dell'autorità giudiziaria.

[32] V. M. BASSINI, *Fundamental rights and private enforcement in the digital age*, in *European Law Journal*, vol. 25, issue n. 2, 2019, p. 187.

[33] Cfr. P. VILLASCHI, *La (non) regolamentazione dei social network e del web*, in M. D'AMICO-C. SICCARDI (a cura di), *La Costituzione non odia: conoscere, prevenire e contrastare l'hate speech online*, Torino, Giappichelli, 2021, p. 113 ss.

[34] Con riferimento a quest'ultima dimensione, tale modalità di intervento è, ad esempio, quella accolta dalla legge n. 71 del 2017 in materia di cyberbullismo (*Disposizioni a tutela dei minori per la prevenzione e il contrasto del fenomeno del cyberbullismo*). Alla base dell'adozione di tale atto vi era la volontà di introdurre una forma peculiare di tutela per i minori, tenendo conto delle caratteristiche peculiari connesse a questa specifica espressione d'odio. Oltre a valorizzare la dimensione preventiva e rieducativa con riferimento a tali reati, infatti, la disciplina prevede la possibilità di presentare, secondo il meccanismo del *notice and takedown*, una richiesta di oscuramento e rimozione di un contenuto alla piattaforma. In forza dell'articolo 2 della legge, ogni adulto o minore ultraquattordicenne possono presentare un reclamo al gestore del sito o del *social media* i quali sono tenuti a esaminare tempestivamente la segnalazione. In caso di mancata notifica della presa in carico di tale istanza o di adozione di un provvedimento entro termini stringenti, è prevista la possibilità di ricorso immediato al Garante per la privacy, V. sul punto L. DAL CORONA, *La legge sul cyberbullismo*, in M. D'AMICO-C. SICCARDI (a cura di), *La Costituzione non odia*, cit., p. 143 ss.; cfr. altresì G. CASSANO, *Stalking, atti persecutori, cyberbullismo e tutela dell'oblio. Aggiornato con la legge 29 maggio 2017*, n. 71, Milano, Wolters-Kluwer, 2017.

incitamento all'odio. Più nel dettaglio, alla luce di tale cornice normativa, il soggetto privato sottoscrivente sarebbe tenuto a: 1) intervenire repentinamente, entro 24 ore, in seguito a una segnalazione, per esaminare - ed eventualmente eliminare – l'elemento illecito; 2) implementare le proprie linee guida per renderle conformi alle previsioni codicistiche e migliorare le procedure interne per la presentazione di reclami[35].

Il *Code of Conduct* è stato adottato il 30 maggio del 2016 ed ha visto l'immediata adesione dei più rilevanti attori della realtà digitale, da Facebook a Google. Sotto il profilo ispiratore, anch'esso nasce dalla volontà di istituire un'alleanza sinergica fra attori privati e istituzioni europee, data la crescente sensibilità di quest'ultime in tale ambito[36]. Difatti, l'elaborazione dell'atto normativo è stata preceduta dal Colloquium "*Tolerance and respect: preventing and combating anti-Semitic and anti-Muslim hatred in Europe*" dell'Internet Forum del dicembre 2015 in cui già la Commissione manifestava un interesse all' individuazione di strumenti adeguati a prevenire la diffusione di discorsi discriminatori con specifico riferimento alle condizioni di alcune minoranze[37], promuovendo un approccio inizialmente fondato, appunto, sul *soft law*[38].

Malgrado il ruolo chiave rivestito dall'Esecutivo europeo, il quale è tenuto ad una valutazione periodica in merito all'applicazione dell'atto[39], sin dalla fase di elaborazione della disciplina pareva emergere la preoccupazione che fossero in realtà le piattaforme aderenti a dover effettuare una valutazione sostanziale dei contenuti, facendosi carico di un delicato bilanciamento fra i valori di rilievo costituzionale in gioco. Nonostante le rassicurazioni provenienti dalle autorità, le quali sottolineavano come, in realtà, i ferrei parametri per la qualificazione dell'oggetto fossero tratti dalla solida giurisprudenza della Corte Europea dei diritti dell'uomo, già nella fase consultiva si prospettava con preoccupazione un'equiparazione delle *Internet platforms* a organi giudicanti chiamati a sindacare la legalità delle condotte espressive[40].

---

[35] Si rimanda al testo originale del Code of Conduct del 2016, https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en.

[36] Cfr. Commission staff working document "*Countering racism and xenophobia in the EU: fostering a society where pluralism, tolerance and non-discrimination prevail*", in Brussels, 15.3.2019, SWD(2019) 110 final.

[37] Cfr. le dichiarazioni della Commissione relative alla pubblicazione del Codice, European Commission, Press Release: *European Commission and IT Companies announce Code of Conduct on illegal online hate speech*, Bruxelles, 31 maggio 2016.

[38] Inoltre, la pubblicazione del Codice è stata seguita dalla emanazione di una Comunicazione contenente importanti linee guida relative all'implementazione del Codice, v. Communication from the Commission to the European Parliament, the Council, the European economic and social committee and the Committee of the regions *Tackling Illegal Content Online - Towards an enhanced responsibility of online platforms*, Bruxelles, 28 settembre 2017, COM (2017) 555.

[39] Sin dal 2016, infatti, l'applicazione del Codice è soggetta ad una valutazione annuale svolta in collaborazione con le organizzazioni di settore collocate sul territorio. Il settimo rapporto è stato pubblicato a novembre 2022, cfr. https://ec.europa.eu/info/sites/default/files/2022_11_21_fs_code_of_conduct.pdf

[40] Date le elevate esigenze di tutela che in questa sfera emergerebbero, tale scenario tenderebbe a configurarsi come una vera e propria "privatizzazione della giustizia". Per un approfondimento concernente il processo partecipativo di formazione del codice e delle attività di analisi e monitoraggio cfr. K. PODSTAWA, *Hybrid Governance or… Nothing? The EU Code of Conduct on Combatting Illegal Hate Speech Online*, in E. CARPANELLI- N. LAZZARINI (a cura di), *Use and Misuse of New Technologies. Contemporary Challenges in International and European Law*, Springer, 2019, p. 171 ss.

Partendo proprio dalle criticità evidenziate, il contributo intende focalizzarsi proprio su quest'ultimo modello normativo invalso a livello euro-unitario per ostacolare la diffusione di discorsi d'odio con lo scopo di porre in luce le principali incrinature che tale modello di *proxy censorship* farebbe trasparire con riferimento ai principi costituzionali e alla tutela dei diritti fondamentali degli utenti. Nell'analizzare la strategia continentale volta al contrasto di tale peculiare forma di illecito si focalizzerà l'attenzione sui seguenti profili che attengono al *Codice di condotta per lottare contro la diffusione dell'incitamento all'odio online*: la fattispecie, lo strumento e il contesto normativo. Più specificamente l'obiettivo del contributo è quello di soffermarsi sui seguenti aspetti: 1)le difficoltà definitorie concernenti l'*hate speech*, una questione già universalmente riscontrata in tale delicata materia che andrebbe a presentarsi anche nella formulazione dell'atto normativo di *soft law*; 2) le problematicità scaturenti dall'utilizzo dello strumento algoritmico il quale, data la complessità dei contenuti esaminati, rischia sovente non solo di incentivare forme di censura collaterale ma anche di colpire proprio categorie minoritarie e soggetti vulnerabili, frustrando proprio quelle finalità di giustizia sostanziale poste alla base dell'intervento normativo; 3) il *framework* normativo generale all'interno del quale il *Code* si inserisce oggi, caratterizzato da una percorso di riforma delle responsabilità dei prestatori di servizi dell'informazione di fronte alla diffusione di contenuti illegali, il cui atto-chiave è rappresentato dal *Digital Service Act*.

Invero, lo scopo della riflessione è quello di analizzare i primi due elementi, connaturati al Codice di condotta sin dalla sua approvazione, alla luce di una cornice regolatoria profondamente rimodellata dalla nuova disciplina. In sintesi, la finalità ultima che il contributo si propone è quella di comprendere se le opacità presenti nel *Code*, le quali risultano decisive nel modellare l'attività censoria delle *Internet Platforms*, si rivelino "superate" all'interno di un quadro garantistico radicalmente innovato dal DSA, una normativa che individua nella lotta ai contenuti illegali e nel rafforzamento delle tutele individuali i suoi presupposti fondativi.

2.1 La vaghezza della fattispecie: la definizione di *hate speech* emergente dal Codice di condotta e il conseguente ampliamento della discrezionalità delle *Internet platforms*

Il primo elemento che è opportuno prender in esame riguarda proprio il fondamento legale posto alla base dell'intervento delle piattaforme. Fin dalla pubblicazione dell'atto, infatti, la dottrina segnalava come la definizione di *hate speech* derivante dalle previsioni codicistiche potesse qualificarsi come "*overly broad*", una genericità tale da estendere eccessivamente la discrezionalità delle *IT companies*, le quali diverrebbero capaci di limitare profondamente la libertà di espressione in presenza di una carente applicazione dei

principi di legalità e proporzionalità[41]. Il codice, infatti, rimanda espressamente alla formulazione contenuta all'interno della decisione quadro del 2008 "*EU Framework Decision on Combating Certain Forms and Expressions of Racism and Xenophobia by Means of Criminal Law*" (Framework Decision 2008/913/GAI)[42] e alle leggi nazionali di attuazione. La stessa, tuttavia, già presentava una qualificazione di discorso d'odio non particolarmente convincente e, per giunta, reputata non conforme agli standard internazionali[43]. Difatti, la disciplina identifica l'*hate speech* come «*incitement to hatred*» e non, più dettagliatamente, come «*incitement to violence, discrimination e hostility*»[44]. Questa qualificazione è stata reputata vaga, fuorviante e giuridicamente non adeguata in quanto collegabile a uno stato emotivo piuttosto che al verificarsi di un pericolo certo. La disciplina eurounitaria, inoltre, pur elencando varie tipologie di reato, predilige un approccio repressivo, valorizzando, per giunta, le sanzioni detentive in palese contrasto con il principio di proporzionalità e gradualità[45].

La questione del corretto inquadramento della condotta lesiva ha sempre costituito, d'altra parte, un aspetto cruciale in questo delicato ambito, una sfida che ha accompagnato la lunga tradizione giuridica sovranazionale e continentale volta a prevenire e contrastare qualsivoglia forma di incitamento all'odio nei confronti di una persona «*on the basis of race, religion, ethnicity or national origin*» [46]. Esistono, infatti, una pluralità di risalenti atti chiave che hanno contributo a edificare una robusta cornice garantistica[47], dalla più generiche formulazioni prescrittive contenute all'interno della Dichiarazione universale dei diritti dell'uomo del 1948 al Patto sui diritti civili e politici del 1966[48], fino ad interventi più articolati, tendenti

[41] B. BUKOVSKA, *The European Commission's Code of Conduct for Countering Illegal Hate Speech Online. An analysis of freedom of expression implications,* in *Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression*, *Article 19 - The Netherlands*, 2019, p. 2.

[42] La decisione quadro viene considerata un atto cardine del processo di edificazione di un quadro normativo europeo conforme ai valori fondanti posti dai Trattati e dalla Carta dei diritti fondamentali e in linea con i principi di tolleranza, pluralismo e non discriminazione. Nella visione della Commissione, l'atto porrebbe «*the framework for a common response to racist hate speech and hate crime. It obliges Member States to penalise the public incitement to violence or hatred against persons defined by reference to race, colour, religion, descent or national or ethnic origin, including when committed online. It also requires them to ensure that the racist and xenophobic motivation is considered as an aggravating circumstance or can be taken into account in the determination of the penalties for any other criminal offence*», in Commission staff working document *Countering racism and xenophobia in the EU*, cit., p. 3.

[43] A. PORTARU, *Freedom of Expression Online: The Code of Conduct on Countering Illegal Hate Speech* Online, in *Revista Romana de Drept European*, n. 4, 2017, pp. 82-83.

[44] B. BUKOVSKA, *The European Commission's Code of Conduct for Countering Illegal Hate Speech Online*, cit., afferma, infatti come «*incitement to hatred makes the proscribed outcome an emotional state or opinion, rather than the imminent and likely risk of a manifested action (discrimination, hostility or violence*)».

[45] In merito alla decisione quadro del 2008 e alle sue luci ed ombre cfr. T. MOSCHETTA, *La decisione quadro 2008/913/GAI contro il razzismo e la xenofobia: una «occasione persa» per l'Italia?*, in *Rivista di Diritto dell'Economia, dei Trasporti e dell'Ambiente*, Vol. XII, 2014, p. 31 ss.

[46] M. ROSENFELD, *Hate Speech in Constitutional Jurisprudence: A Comparative Analysis*, in *Cardozo Law Review*, n. 24, 2003, p. 1523.

[47] P. DE SENA, M. CASTELLANETA, *La libertà di espressione e le norme internazionali*, cit.

[48] La Dichiarazione si limita a ribadire il divieto di non discriminazione come conseguenza della cogenza del principio di uguaglianza (art. 7) mentre il Patto del 1966 introduce già una definizione più articolata la quale esclude la tollerabilità di qualsivoglia forma di «incitamento alla discriminazione, all'ostilità o alla violenza» (art. 20).

non solo a introdurre forme di tutela specifiche a favore di singoli gruppi ma nei quali inizia ad esser visibile anche un minimo sforzo definitorio, quali, ad esempio, la Convenzione per l'eliminazione di ogni forma di discriminazione razziale del 1965[49] e la Convenzione per l'eliminazione di ogni forma di discriminazione contro le donne[50].

La fioritura di fonti convenzionali in materia trova certamente alla base un solido presupposto storico-ideologico: un fermo rifiuto verso qualsivoglia tolleranza nei confronti di manifestazioni inneggianti alla violenza aventi radici discriminatorie, a seguito della tragica esperienza consumatasi fra le due guerre, un fondamento unificante che ha contribuito a modellare lo stesso approccio continentale alla spinosa questione dei limiti da porre alla libertà di espressione. Quest'ultimo, infatti, non ritiene incompatibile con i valori essenziali dell'ordinamento democratico l'apposizione di limitazioni alla manifestazione del pensiero qualora esse siano finalizzate a salvaguardare beni di rilievo costituzionale, un'ottica normativa icasticamente rappresentata dalla stessa formulazione dell'articolo 10 par. 2 della CEDU e poi sviluppata dalla giurisprudenza della Corte di Strasburgo. Il giudice sovranazionale, invero, ha svolto una funzione determinante nel rendere intellegibile la portata garantistica della disposizione convenzionale stabilendo, appunto, come l'*hate speech* si sostanzi in sé in una condotta incompatibile con il sistema di tutele posto dalla Carta e come le misure legislative finalizzate a contrastare discorsi violenti possano costituire limitazioni legittime della libertà di espressione in favore della tutela necessaria della reputazione degli individui e delle libertà fondamentali[51]. Esso ha altresì esercitato un ruolo cruciale nel dilatare i confini della garanzia convenzionale, andando a ricondurre sotto la definizione di «espressioni che incitano, promuovono, giustificano l'odio fondato sull'intolleranza»[52] una pluralità di contenuti discriminatori, anche quelli eccedenti le più tradizionali fattispecie di istigazione all'odio razziale e religioso, quali, ad esempio, manifestazioni omofobiche[53] e dichiarazioni negazioniste[54].

[49] L'articolo 4 della Convenzione, infatti, oltre a condannare la commissione di questa tipologia di violazioni, le definisce più chiaramente inquadrandole come condotte che «incoraggino» o «giustifichino» ogni forma di odio e di discriminazione razziale.

[50] Cfr. M. D'AMICO, *Audizione nell'ambito di un'indagine conoscitiva sulla natura, cause e sviluppi recenti del fenomeno dei discorsi d'odio, con particolare attenzione alle evoluzioni della normativa europea in materia*, 20 luglio 2021, disponibile in www.senato.it.

[51] V. Corte europea dei diritti dell'uomo, I. sez., *Gündüz contro Turchia*, 14 giugno del 2006. Analoghe motivazioni emergono anche in *Soulas e altri contro Francia* (10 luglio 2008), *Féret contro Belgio* (16 luglio 2009).

[52] Corte europea dei diritti dell'uomo, *Erbakan contro Turchia*, 6 luglio 2006.

[53] Cfr. Corte europea dei diritti dell'uomo, *Lìellendahl contro Islanda*, 11 giugno 2018, nella quale la Corte ha occasione di chiarire più nel dettaglio come l'*hate speech* determini violazioni differenti, distinguendo tra «*the gravest forms of hate speech*» e «*the less grave*», fra cui ricaderebbe il contenuto omofobico alla base della condanna della corte nazionale. Il Giudice argomenta, infatti, che «*although the comments were highly prejudicial, as discussed further below, it is not immediately clear that they aimed at inciting violence and hatred or destroying the rights and freedoms protected by the Convention*» (§26). Sulla medesima forma di odio fondato su una discriminazione in base all'orientamento sessuale si vedano altresì *Vejdeland contro Svezia,* 9 febbraio 2012 e *Beizaras and Levickas contro Lituania*, 14 gennaio 2020.

[54] Cfr., fra le pronunce più recenti, *Willimson c. Germania*, 31 gennaio 2019, e *Pastörs c. Germania*, 3 ottobre 2019.

Nonostante ciò, tuttavia, data la rilevanza che tale libertà comunque riveste all'interno del ventaglio valoriale europeo, la corretta e puntuale circoscrizione del comportamento lesivo è rimasta al centro delle preoccupazioni dei legislatori, della dottrina e della giurisprudenza[55]. Da un lato, infatti, è stato chiarito come, affinché si possa parlare di *hate speech*, sia necessaria la presenza di un'espressione «*designed to promote hatred*»[56]. Più nel dettaglio, quindi, non è sufficiente la diffusione di un contenuto verbalmente odioso ma la compresenza di vari fattori: un elemento soggettivo, ovvero una manifesta volontà istigatoria, e un elemento oggettivo[57], ossia l'idoneità del contenuto a determinare un conseguente comportamento violento e il rischio concreto che tale evento di verifichi[58]. Anche in merito a tali aspetti è imprescindibile, ancora una volta, il ruolo chiave esercitato dal giudice di Strasburgo il quale, nonostante l'equilibrato *modus decidendi* volto a evitare che si profilasse una tutela assoluta della libertà di espressione in presenza di altri valori di innegabile rilevanza convenzionale, ha comunque coerentemente ribadito come tali valutazioni non si possano sottrarre ad un rigido scrutinio di proporzionalità[59]. È necessario sottolineare, tuttavia, come discorsi d'odio "di elevata gravità" sono sovente divenuti oggetto di quel differente e più rigido approccio della Corte EDU basato sull'applicazione della *abuse clause* posta dall'articolo 17 della Convenzione. In altre parole, dunque, opinioni giustificazioniste così come forme di propaganda volte ad istigare esplicitamente all'odio nei confronti di determinate minoranze sono state spesso considerate insuscettibili di per sé di beneficiare della tutela posta dall'articolo 10 perché reputate per loro natura

[55] Cfr. M. N. CAMPAGNOLI, *Social media e* information disorder: *questioni di ecologia comunicativa in Rete (Parte seconda - L'hate speech)*, in *Dirittifondamentali.it*, 2/2020, p. 1600 ss.
[56] M. ROSENFELD, *Hate Speech in Constitutional Jurisprudence*, cit., p. 1523.
[57] V. A. BROWN, *What is Hate Speech? Part 1: The myth of hate,* in *Law and Philosophy*, 36, 2017, p. 437.
[58] Cfr. M. D'AMICO, *Odio on line: limiti costituzionali e sovranazionali*, in M. D'AMICO-C. SICCARDI (a cura di), *La Costituzione non odia: conoscere, prevenire e contrastare l'hate speech online*, Torino, Giappichelli, 2021, p. 16; G. ZICCARDI, *Odio online. Violenza verbale e ossessioni in rete*, Raffaello Cortina, Milano, 2016, p. 19 afferma che i faticosi tentativi di chiarimento del significato di *hate speech* confluiscono nella definizione di tre motivi posti alla base della discriminazione - quali nazionalismo, razzismo e il fattore religioso - e in tre tipologie di condotta: discriminazione, ostilità e violenza. Ad ogni modo, è opportuno ricordare che la descrizione elaborata a livello continentale richiami, a tutta evidenza, la nota dottrina del «*clear and present danger*» elaborata dalla Corte suprema statunitense a partire dalle note *dissenting opinion* del giudice Holmes nel caso *Schenk v. United States* del 1919 e del giudice Brandeis nel caso *Abrams v. United States*. Oreste Pollicino osserva come, all'interno del sistema americano nel quale vigerebbe «una tutela sacrale del diritto di parola», in questa definizione di un chiaro limite alla libertà di espressione si annidi in realtà una volontà di consolidare l'ampiezza dello spazio entro il quale la libertà di manifestazione del pensiero solidamente si manifesta, circoscrivendola, per sottrazione, attraverso l'individuazione di tale puntuale limitazione. In altre parole, tale diritto andrebbe ad arrestarsi solo in presenza di una chiara e concreta minaccia per le altre libertà e l'ordine democratico. Peraltro, tale limitazione è stata poi ulteriormente definita nella sentenza *Brandeburg v. Ohio* nella quale i margini della libera manifestazione del pensiero in presenza di discorsi d'odio sono stati ulteriormente dilatati in seguito alla definizione di rigidi parametri per qualificare un discorso come pericoloso per l'ordine pubblico, in O. POLLICINO, *La prospettiva costituzionale sulla libertà di espressione nell'era di Internet,* in G. PITRUZZELLA-O. POLLICINO-S. QUINTARELLI (a cura di), *Parole e potere. Libertà d'espressione hate speech e fake news*, Milano, Egea, 2017, p. 23 ss.
[59] Si veda, ancora una volta, *Erbakan contro Turchia*, 6 luglio 2006, § 56 nella quale la corte ribadisce come «a*ny formalities, conditions, restrictions or penalties imposed are proportionate to the legitimate aim pursued*», nonché il percorso argomentativo intrapreso dalla Corte nella più recente sentenza *Altıntaş contro Turchia*, 10 marzo 2020.

idonee *ab origine* a porsi in contrasto con i valori fondanti della Convenzione[60]. In tali circostanze, dunque, il collegio ha proceduto ad un inasprimento del proprio *modus argomentandi*, arrestando il suo sindacato all'esame del contenuto espressivo: l'effetto distruttivo, a danno delle fondamenta valoriali della Carta, generato da determinate manifestazioni in tali casi limite, legittimerebbe, ad avviso del giudice, un sindacato *content based*, svolto in assenza di quella delicata ponderazione di interessi a cui si ricorre ogni qual volta venga in rilievo l'articolo 10. Tale lettura alternativa, modellata sulla peculiarità dell'espressione d'odio esaminata, ha, ad avviso di alcuna dottrina, contribuito a creare delle discrepanze nella strategia giurisprudenziale dell'organo giurisdizionale volta a inquadrare l'espressione d'odio secondo il modello inaugurato in *Handyside* il quale stabilisce che ogni opinione, anche quella più "scomoda", debba in astratto considerarsi coperta dalla tutela convenzionale e che il livello di garanzia accordato non si possa fondare sulla tipologia del contenuto quanto su una oculata operazione di bilanciamento operata caso per caso che tenga conto delle circostanze normative e fattuali della vicenda[61].

Ad ogni modo, al di là di tali oscillazioni, le finalità espansive sottese alla fonte convenzionale e la puntuale azione definitoria elaborata in seno al Consiglio d'Europa emergono coerentemente anche da numerosi atti *soft law* emanati dal Consiglio dei ministri per definire e reprimere i crimini d'odio, dalla Raccomandazione n. 20 del 1997[62] fino alla recentissima Raccomandazione n. 16 del 2022. Quest'ultima, all'interno del Preambolo, ribadisce l'importanza di una corretta e omogenea definizione di "*hate speech*" a livello continentale, nonché la necessità di introdurre misure di contrasto che risultino «*appropriate and proportionate to the level of severity of its expression*»[63].

Sebbene qualsivoglia parallelismo esiga cautela, data la profonda differenza ontologica intercorrente fra le istituzioni esaminate e gli atti da esse adottati, è da osservare come un simile rigore definitorio sembra tuttavia mancare all'interno del panorama normativo euro-unitario[64]. Come già accennato, infatti, il *Code of Conduct for Countering Illegal Hate Speech Online* è stato elaborato in un clima istituzionale connotato da un crescente interesse delle istituzioni europee per l'innalzamento del livello di tutela e ispirato dalla volontà di conferire maggior effettività all'articolo 21 della Carta di Nizza. Negli ultimi anni, inoltre, le autorità

[60] Si tratta di un filone inaugurato nella decisione *Glimmerveen and Hagenbeek v. the Netherlands* del 1979, poi ripreso nella decisione *Kühnen v. Federal Republic of Germany* del 1986. Si veda l'applicazione della clausola anche nella nota sentenza concernente espressioni negazioniste *Garaudy v. France* del 2003 fino alla sua estensione a manifestazioni islamofobiche in *Norwood v. United Kingdom* del 2004.

[61] Sul punto cfr. l'analisi di C. Caruso, il quale, oltre ad argomentare criticamente sulle modalità di applicazione dell'articolo 17 e sull' «effetto ghigliottina» derivante da tale approccio, sottolinea come la scelta di propendere per l'una o altra strategia argomentativa non risulti sempre intellegibile, v. C. CARUSO, *L'hate speech a Strasburgo: il pluralismo militante del sistema convenzionale*, in *Quaderni costituzionali*, fasc. 4/2017, p. 963 ss.

[62] *Recommendation no. R (97)20 of the Committee of Ministers to member states on "hate speech"*, disponibile a https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=0900001680505d5b.

[63] *Recommendation CM/Rec(2022)16 of the Committee of Ministers to member States on combating hate speech*, disponibile a https://search.coe.int/cm/Pages/result_details.aspx?ObjectID=0900001680a67955#_ftn1.

[64] V. NARDI, *I discorsi d'odio nell'era digitale: quale ruolo per l'internet service provider*? in *Penalecontemporaneo.it*, 2019, p. 8 ss.

sembrano aver compiuto ulteriori passi avanti: ridurre e contrastare l'odio *online* è divenuto infatti uno degli obiettivi chiave della Presidente Von der Leyen posto alla base dell'*European Democracy Action Plan* del 2020 (EDAP)[65]. Inoltre, proprio in forza di tali presupposti, la Commissione aveva proposto una modifica dell'articolo 83 par. 1 del TFUE, il quale stabilisce che, per alcuni reati di particolare gravità aventi il carattere della transnazionalità, Parlamento e Consiglio possano definire delle direttive per precisare le fattispecie e dare indicazioni minime. L'obiettivo perseguito dall'Unione consiste proprio nell'introduzione all'interno di tale categoria di tutte le forme di *hate crime* e *hate speech*[66]. Tale modifica consentirebbe un'estensione delle fattispecie a qualsivoglia condotta discriminatoria consumata online, persino quelle *gender-based*[67], la cui eliminazione rientra peraltro fra gli obiettivi stabiliti nella EU Gender Equality Strategy 2020-2025[68].

La vaghezza definitoria emergente dal *Code of Conduct for Countering Illegal Hate Speech Online* dovrebbe forse esser inquadrata e compresa proprio alla luce di tale clima riformatore ispirato da velleità di espansione delle tutele poste a salvaguardia del principio di non discriminazione in chiave universalizzante e non escludente[69]. Più chiaramente, forse, potrebbe essere immaginata come un effetto negativo collaterale prodotto dal sopradescritto ambizioso progetto europeo, un percorso animato da finalità meritevoli ma forse non in grado di ponderare oculatamente come, in tal caso, l'assenza di determinatezza e tassatività della disciplina avrebbe l'effetto di dilatare eccessivamente i margini interpretativi delle piattaforme private[70], per giunta con profonde discrepanze applicative fra gli aderenti, creando un'ulteriore frammentazione e differenziazione delle garanzie[71].

---

65 L'introduzione di meccanismi adeguati e lo svolgimento di sforzi ulteriori appare necessario per il rafforzamento delle istituzioni democratiche in un'ottica sostanziale. La presenza di fenomeni di incitamento all'odio e la diffusione di contenuti odiosi viene considerata uno dei fattori pregiudizievoli che ostacola l'espressione di idee da parte delle minoranze, cfr. *Communication from the Commission to the European Parliament, the Council, the European economic and social Committee and the Committee of the regions on the European democracy action plan*, 3 dicembre 2020, par. 2.4.

66 Difatti, secondo l'articolo 83 par. 1, il Parlamento ed il Consiglio possono dettare «norme minime relative alla definizione dei reati e delle sanzioni in sfere di criminalità particolarmente grave che presentano una dimensione transnazionale derivante dal carattere o dalle implicazioni di tali reati o da una particolare necessità di combatterli su basi comuni». Benché tali crimini abbiano carattere tassativo, il Consiglio dispone della possibilità di estendere la lista di reati deliberando all'unanimità, previo parere del Parlamento, sul punto cfr. N. PERSAK, *Criminalising hate crime and hate speech at EU level: extending the list of eurocrimes under article 83(1) TFEU*, in *Criminal Law Forum*, 2022, p. 85 ss.

67 Con riferimento a tale categoria di crimini d'odio, la *European Commission against racism and intolerance* (ECRI), operante in seno al Consiglio d'Europa, segnalava un verificarsi crescente di tali manifestazioni d'odio durante tutto il 2020, cfr. https://rm.coe.int/annual-report-on-ecri-sactivities-for-2020/1680a1cd59.

68 In merito ai punti chiave di tale strategia si rimanda al testo della comunicazione *Communication from the Commission to the European Parliament, the Council, the European economic and social Committee and the Committee of the Regions A Union of Equality: Gender Equality Strategy 2020-2025*, COM/2020/152.

69 Si veda la ricostruzione di O. POLLICINO – G. DE GREGORIO, *Hate speech: una prospettiva di diritto costituzionale comparato*, in *Giornale di diritto amministrativo*, 4/2019 p. 429 ss.

70 V. F. ABBONDANTE, *Il ruolo dei social network nella lotta all'hate speech: un'analisi comparata fra l'esperienza statunitense e quella europea*, in *Informatica e Diritto*, 2017, p. 41 ss.

71 F. CASAROSA, *L'approccio normativo europeo verso il discorso dell'odio online: l'equilibrio fra un sistema di "enforcement" efficiente ed efficace e la tutela della libertà di espressione*, in *Questionegiustizia.it*, 2020.

2.2 La "fallibilità" dello strumento algoritmico e l'impatto sui diritti fondamentali degli utenti e delle categorie vulnerabili

L'ampia formulazione della previsione presente *ab origine* costituisce, dunque, una criticità suscettibile di minare gli obiettivi garantistici posti alla base dell'intervento normativo, un fattore in grado di contribuire alla compressione della libertà di manifestazione del pensiero degli utenti nello spazio virtuale. Questo effetto involontario scaturente dalla disciplina sembra ulteriormente aggravato dall'utilizzo di sistemi di intelligenza artificiale finalizzati alla rilevazione di contenuti discriminatori qualificabili come *hate speech*. Com'è noto, il ricorso a meccanismi automatizzati nell'ambito dell'attività di moderazione costituisce un compromesso oggigiorno ineliminabile per gli intermediari[72]. Dietro tale soluzione, determinante in un sistema caratterizzato dall'interazione di milioni di contenuti, si cela altrettanta consapevolezza circa i rischi generali connessi al loro uso[73].

Non è certo un dato sconosciuto il fatto che tutti i software di IA siano soggetti a un certo tasso di errore, dato il generico funzionamento su basi probabilistiche[74]. Tali problematiche, tuttavia, si rivelano cruciali in presenza di contenuti d'odio[75], data la ricorrente pre-esistenza, largamente denunciata in dottrina, di *bias* involontari[76]. Il margine di fallibilità, dunque, sembra elevarsi quando l'oggetto esaminato costituisce un argomento testuale complesso e non immediato nel suo significato. Sebbene si assista, oggi, a tentativi di perfezionamento del dispositivo sotto il profilo della corrispondenza fra strumento adoperato e tipologia di oggetto – verbale, audiovisivo, misto – e, più specificatamente, l'utilizzo di meccanismi di *natural language processing* sempre più sofisticati[77], ancora permane una accertata incapacità per la macchina di comprendere il metatesto dell'espressione esaminata. Elementi intrinseci dell'oggetto analizzato quali, ad esempio, sfumature semantiche[78], ovvero l'uso ironico o satirico di una determinata terminologia,

---

[72] V. T. GILLESPIE, *Content moderation, AI, and the question of scale*, in *Big Data and Society*, 2020.

[73] Cfr. O. POLLICINO, *Judicial protection of fundamental rights on the Internet*, Oxford, Hart, 2021, pp. 193-194.

[74] G. DE GREGORIO–O. POLLICINO- P. DUNN, *Digitisation and the central role of intermediaries in a post-pandemic world*, in *Medialaws.eu*, 2021.

[75] R. GORWA-R. BINNS-C. KATZENBACH, *Algorithmic content moderation: Technical and political challenges in the automation of platform governance*, in *Big Data & Society*, vol. 7, 2020, p. 1 ss.

[76] G. ZICCARDI, *Odio online. Violenza verbale e ossessioni in rete*, Raffaello Cortina, Milano, 2016, p. 97 ss.

[77] E. LLANSÓ et AL., *Artificial intelligence, Content Moderation, and Freedom of Expression*, Transatlantic working group, 26 February 2020, p. 5. Gli autori si soffermano su un classico esempio di Sistema di NLP, il Sistema di IA chiamato "Perspective", strumento altamente sofisticato e congegnato per identificare la presenza di espressioni tossiche, evidenziandone luce ed ombre: «*Perspective provides a good illustration of both the capabilities and limitations of a sophisticated NLP tool. It has been used for a variety of applications, including as a tool that comment moderation systems use to warn users that they may be posting a "toxic" comment and to give them the opportunity to revise their comment. But Perspective, and the concept of evaluating "toxicity" of comments, is far from perfect; soon after the Perspective API was launched, researchers began exploring ways to "deceive" the tool and express negativity that slipped under the radar, and researchers continue to identify bias in the tool, such as misclassification that disproportionately affects different racial groups*».

[78] Tale "insensibilità" dell'algoritmo è stata oggetto di numerosi studi, in tempi in cui ancora non si faceva largo uso dei sistemi di IA, finalizzati ad analizzare la capacità del sistema di rilevare correttamente *tweet* razzisti nei confronti della comunità afroamericana, cfr., ad esempio, I. KWOK-Y. WANG, *Locate the Hate: Detecting Tweets against Blacks*, in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 27, 2013.

sfuggirebbero spesso allo *screening* operato dal sistema[79]. In altre parole, nell'ambiente virtuale si andrebbe ad esacerbare l'annoso problema della "contestualizzazione" dei discorsi odiosi e discriminatori, già ampiamente osservato nella dimensione *offline*[80]. Dunque, malgrado l'esistenza di sforzi multisettoriali mossi dallo scopo di realizzare un "*debiasing*" dei sistemi[81], ancora resistono numerose disfunzionalità connaturate allo strumento in grado di produrre un ampio numero di "falsi positivi"[82], una distorsione che si rivela direttamente proporzionale alla rigidità dei limiti posti dalla disciplina, cioè alla severità della stessa che, come evidenziato sinora, nel caso del *Code of Conduct* è massima.

La problematica indagata, se vagliata attraverso la lente della tutela dei diritti fondamentali dell'utente, induce a delineare uno scenario particolarmente opaco. Difatti, la descritta alterazione acquisisce particolare rilevanza all'interno di tale analisi poiché non solo costituisce un aspetto applicativo chiave della lotta all'*hate speech* che, osservato unitamente alla precaria base legislativa, contribuisce a espandere la funzione censoria delle IT; ma, altresì, poiché l'effetto negativo derivante dal suo utilizzo sembra colpire in particolare proprio quei gruppi minoritari e soggetti vulnerabili ai quali le finalità garantistiche dell'intervento normativo sembrano guardare[83]. Studi a carattere multidisciplinare condotti su basi quantitative constaterebbero infatti come il sistema classifichi più frequentemente come "offensivo" - e, dunque, come contenuto da rimuovere - l'uso che di un determinato linguaggio viene fatto da alcune minoranze, spesso con una mera finalità descrittiva, in chiave solidaristica o, addirittura, a scopo performativo, cioè con la esplicita e ambiziosa finalità di "epurare" la terminologia utilizzata dal suo significato originariamente "odioso", allo scopo di trasformarlo in un baluardo lessicale identitario[84].

---

[79] P. DUNN, *Moderazione automatizzata e discriminazione algoritmica: il caso dell'hate speech, in Rivista italiana di informatica e diritto*, 1/2022, p. 135 ss.

[80] A. WILSON-M. LAND, *Hate Speech on Social Media: Content Moderation in Context*, in *Connecticut Law Review*, vol. 52, 2021, n. 3, affermano come la falla principale della moderazione online dei discorsi odiosi risieda proprio nel fatto che essi risultano «*deeply acontextual*». Utilizzando gli studi condotti da Lawrence nella dimensione *offline* volti a valorizzare l'importanza fra *hate speech* e contesto, nonché l'importanza della relazione intercorrente fra interlocutore e "vittima" potenziale della manifestazione d'odio - fattore non sempre trasferibile nella dimensione virtuale in cui il soggetto spesso si rivolge ad una platea indefinita - gli autori concludono che «*social media prohibitions on hate speech, in contrast, prohibit speech based on content rather than context. There is no consideration of local context, country- or region-specific meanings, the identity of the speaker or the target, or the relationship between speaker and listener. (…) This deliberate disregard of power and context leads to overbroad and even irrational results*» (p. 1058 ss.).

[81] È sconfinata, infatti, la ricerca volta al temperamento dei *bias* più diffusi quali, ad esempio, i *lexical bias,* il cui impatto si tenta di mitigare mediante l'aggiunta di dati di raffronto che coadiuvano il sistema di IA nell'inquadrare la parola all'interno del corretto contesto espressivo, v. L. DIXON ET. AL., *Measuring and Mitigating Unintended Bias in Text Classification*, in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, Dicembre 2018, p. 67 ss.

[82] Cfr. S. QUINTARELLI, *Content moderation: i rimedi tecnici*, in G. PITRUZZELLA-O. POLLICINO-S. QUINTARELLI (a cura di), *Parole e potere*, cit., p. 112.

[83] P. DUNN, *Moderazione automatizzata e discriminazione algoritmica*, cit., p. 137 osserva come, in realtà, tale effetto distorsivo, peraltro, si rilevi anche a livello di *content curation* e non solo nell'*hard moderation* in senso stretto. Più specificamente, è stato sottolineato in dottrina come il meccanismo algoritmico tenda a perseguire una generale massimizzazione dell'*engagement* dei gruppi di maggioranza con l'effetto di ridurre sempre più percettibilmente gli spazi riservati a gruppi marginalizzati.

[84] L. DIXON ET. AL., *Measuring and Mitigating Unintended Bias*, cit., p. 67.

Questa tendenza, registrata in relazione a diversi gruppi minoritari quali, ad esempio, la comunità LGBTQ+[85] e la comunità afroamericana[86], si scontrerebbe, dunque, con il meccanismo di funzionamento dell'algoritmo[87]. Difatti, le principali disfunzionalità che ancora si rilevano si sostanziano essenzialmente in *lexical bias* e *dialectal bias*[88]. I primi derivano dalla meccanica associazione di una parola a un contenuto "tossico" mentre gli ultimi derivano dall'incapacità del sistema di cogliere l'uso peculiare che si fa di un determinato lessico all'interno di una subcategoria linguistica o dialetto, quale, ad esempio, l'*African American English*.

Guardando a tali dati nel loro complesso, emerge come l'approccio normativo dominante sotteso al funzionamento dell'algoritmo prediliga una logica ispirata a un criterio di uguaglianza formale, ovvero soluzioni proiettate più frequentemente verso l'eliminazione dei cosiddetti "*technical bias*" e a modalità "non interventiste", basate cioè su "*bias preserving" fairness metrics*, cioè tese al mantenimento dello *status quo* e degli equilibri sociali pre-esistenti, con lo scopo di non creare maggiori disuguaglianze[89]. Le conseguenze che ne derivano - come nel caso di specie portato ad esempio - sono, tuttavia, distorsive, poiché tale struttura operativa determina un effetto censorio di gran lunga superiore per alcune minoranze, dando luogo, ad un secondario effetto discriminatorio a base razziale o fondato sull'orientamento sessuale[90]. In ultima istanza, dunque, in assenza di adeguati correttivi tecnici applicati al sistema di IA che tengano conto delle sfaccettate specificità connaturate al linguaggio utilizzato, l'effetto paradossale che l'utilizzo di tale strumento può generare consiste proprio in un rovesciamento del principio di uguaglianza

[85] Si tratta di risultati che, ovviamente, fotografano tendenze emerse prima nelle interazioni reali e solo successivamente virtuali. A tal proposito, si rimanda alle rilevanti riflessioni di natura linguistica e sociologica scaturite dallo studio sulla cosiddetta "*mock impoliteness*" all'interno della comunità LGBTQ+ e, in particolare, la piccola comunità delle *drag queen*, dove l'uso del linguaggio apparentemente offensivo assume una funzione volta al rafforzamento dei legami interni fra i consociati ma anche uno scopo di edificare una sub-cultura, v. S. MCKINNON, *"Building a thick skin for each other". The use of 'reading' as an interactional practice of mock impoliteness in drag queen backstage talk*, in *Journal of Language and Sexuality*, vol. 6, 2017, n. 1, p. 95.

[86] SAP et AL., *The Risk of Racial Bias in Hate Speech Detection*, in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, p. 1168 ss.

[87] Cfr. T. DAVIDSON ET AL., *Automated Hate Speech Detection and the Problem of Offensive Language*, in *arXiv: Computation and Language*, 2017.

[88] ZHOU ET. AL., *Challenges in Automated Debiasing for Toxic Language Detection*, in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2021, p. 3143 ss.

[89] Si vedano le riflessione e le proposte correttive, basate sui meccanismi di *bias transfroming*, di S. WATCHER ET AL., *Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law*, in *Virginia Law Review*, vol. 123, issue 3, 2021, p. 735 ss. i quali argomentano come, quest'ottica correttiva, ispirata a un concetto di uguaglianza sostanziale, costituirebbe comunque un approccio che si concretizza in una misura di discriminazione positiva volta a temperare disparità sociali, un intervento che necessiterebbe quindi di una solida giustificazione legislativa.

[90] Alcune analisi statistiche sembrano infatti dimostrare come l'incapacità dello strumento di moderazione automatizzata di cogliere il contesto porti ad un maggior numero di contenuti eliminati di utenti appartenenti alla comunità delle *drag queen* rispetto a contenuti discriminatori di nazionalisti bianchi, cfr. DIAS OLIVA ET AL., *Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online*, in *Sexuality and Culture*, vol. 25, 2021, p. 700 ss.

sostanziale sottesa al percorso normativo euro-unitario, modellato sull'accrescimento delle tutele e volto all'eliminazione di ogni forma di disparità di trattamento.

3. Il Codice alla prova del contesto: le potenzialità e i limiti del nuovo *framework* garantistico introdotto dal *Digital Services Act*

Alla luce del quadro sinora tracciato, si può dunque affermare che la disciplina introdotta dal Codice, volta a ridurre e contrastare le manifestazioni d'odio *online*, presenti, sia sul versante definitorio sia sul fronte applicativo, dei fattori intrinseci poco convincenti e suscettibili di esercitare un impatto significativo sui diritti fondamentali degli utenti.

Queste "falle originarie" riscontrate nell'atto sembrerebbero oggi assumere un nuovo significato se calate all'interno del contesto normativo odierno. Difatti, nell'attuale momento storico, il *Code of Conduct* non può che essere valutato se non come un tassello di una cornice regolatoria soggetta ad un percorso di radicale trasformazione che trova nel *Digital Markets Act* e nel *Digital Services Act* i suoi atti chiave[91]. Tali regolamenti, infatti, recentemente adottati dopo un lungo *iter* e un faticoso confronto istituzionale[92], costituirebbero gli strumenti operativi per affermare la cosiddetta "sovranità digitale o tecnologica dell'Unione"[93]. Con quest'ultima espressione si identifica l'aspirazione all'edificazione di un'azione strategica autonoma in tale ambito attraverso la creazione di meccanismi "difensivi" e "offensivi" in un settore in cui l'influenza economica e sociale delle *Big Tech* acquisisce risvolti preoccupanti[94] poiché pare accrescersi sempre di più al di fuori dei vincoli imposti dalla normativa euro-unitaria e dai principi

[91] Si veda la risoluzione del Parlamento europeo del 20 ottobre del 2020 "*Risoluzione del Parlamento europeo del 20 ottobre 2020 recante raccomandazioni alla Commissione sulla legge sui servizi digitali: migliorare il funzionamento del mercato unico (2020/2018(INL)*", in https://eur-lex.europa.eu/legal-content/IT/TXT/?uri=CELEX:52020IP0272.

[92] La proposta di regolamento è stata presentata il 15 dicembre 2020. In seguito, il regolamento è stato approvato il 5 luglio 2022 dal Parlamento e successivamente dal Consiglio in data 4 ottobre 2022. Il 27 ottobre 2022 è stato pubblicato sulla Gazzetta Ufficiale dell'Unione. Per una lettura onnicomprensiva del provvedimento e delle diverse fasi di elaborazione cfr. S.F. SCHWEMER, *Digital Services Act: A Reform of the e-Commerce Directive and Much More*, prepared for A Savin, Research Handbook on EU Internet Law, October 2022, p. 1 ss.

[93] In merito a tale concetto e ai suoi possibili significati cfr. L. FLORIDI, *The Fight for Digital Sovereignty: What It Is, and Why It Matters, Especially for the EU*, in *Philosophy and Technology*, 2020, il quale afferma che l'edificazione della «*digital sovereignty is not about replacing national modern-analogue sovereignty, which is necessary but increasingly insufficient. It is about complementing it with a supranational, contemporary-digital one—which is often its condition of possibility—also to provide to all actors and stake-holders wider benefits of harmonization*» (p. 375). Sulle ambiguità connesse all'idea di "sovranità digitale" cfr. V. BERTOLA, *La sovranità digitale e il futuro di Internet*, in *Rivista italiana di informatica e diritto*, fasc. 1/2022, p. 39 ss. Sulla scivolosità del termine, anche sotto il profilo linguistico, v. S. COUTOURE- S. TOUPIN, *What does the notion of "sovereignty" mean when referring to the digital?*, in *New Media and Society*, vol. 21, issue 10, 2018, p. 2305 ss.

[94] Il richiamo alla sovranità digitale è diventato cruciale all'interno del panorama europeo fra il 2019 e il 2020. Nella visione del Parlamento europeo essa si sostanzia nel semplice obiettivo di rendere l'Unione un soggetto «capace di agire autonomamente nella dimensione digitale», come emerso nella scheda esplicativa "*Digital sovereignty for Europe*" redatta dall'organo assembleare nel luglio 2020. Il primo tassello della "strategia digitale" dell'Unione è stato la proposta di regolamento europeo in materia di governance dei dati, il *Data governance Act*, presentata il 25 novembre 2020, il cui dichiarato obiettivo era quello di assicurare una corretta gestione dei dati utilizzabili con il fine di rafforzare la fiducia negli intermediari e di potenziarne i meccanismi di condivisione in tutta l'UE.

fondamentali dell'ordinamento[95]. Utilizzando quest'ultimi come faro guida, il *Digital Service Act*, in particolare, si focalizza proprio sulla regolamentazione dell'offerta dei servizi digitali, promuovendo una disciplina solidamente basata su tre presupposti cardine: trasparenza, responsabilità e garanzia nei confronti dell'utente[96]. In realtà, il dichiarato obiettivo perseguito dalla riforma è ad ampio raggio in quanto aspira, in particolare, ad un efficientamento in senso "democratico" del mercato unico dei servizi di intermediazione, mediante la ricerca di un corretto contemperamento con la tutela dei diritti[97], al fine di favorire la creazione di un «*safe, predictable and trusted online environment*»[98]. A ben vedere, quindi, il legislatore europeo pare aver tradotto in via normativa la presa di coscienza del fatto che il mutamento radicale della configurazione dei *provider* costituirebbe un nuovo fattore di rischio per i cittadini, aggravato dall'assenza di un adeguato servizio di vigilanza e coordinamento di tipo amministrativo, e da un sistema normativo non armonizzato[99]. In particolare, la radicale e complessa riconfigurazione intercorsa negli ultimi vent'anni del servizio di *hosting*, il quale non si limiterebbe più ad operare come contenitore neutrale di contenuti, avrebbe reso quanto mai necessaria la modifica della disciplina previgente regolante la responsabilità dei *provider*, quella posta dall'obsolescente direttiva 2000/31/CE[100].

Il principio guida posto da quest'ultima, la regola tendenziale della irresponsabilità del prestatore per i contenuti da esso stesso ospitati, permane[101], ma viene riconfigurato alla luce delle nuove sfide poste da uno scenario digitale mutato[102], mediante l'introduzione di modifiche puntuali concernenti in particolare

95 Tali obiettivi sono poi confluiti nella Comunicazione "*Plasmare il futuro digitale*" nella quale si evidenzia come, affinché l'UE possa diventare un attore forte e indipendente nell'ambito tecnologico, sia necessario un solido quadro regolatorio che garantisca un'affidabile interazione fra società e persone. Più specificamente, l'art. 2 par. 2 della Comunicazione chiarisce come «la sovranità tecnologica europea inizia dal garantire l'integrità e la resilienza dell'infrastruttura di dati, delle reti e delle comunicazioni e richiede la creazione delle giuste condizioni affinché l'Europa possa sviluppare e utilizzare le proprie capacità chiave, riducendo in tal modo la dipendenza da altre parti del mondo per le tecnologie più importanti. Tali capacità rafforzeranno l'abilità dell'Europa nel definire le proprie regole e i propri valori nell'era digitale, cfr. Comunicazione della Commissione al Parlamento europeo, al Consiglio, al Comitato economico e sociale europeo e al Comitato delle regioni "*Plasmare il futuro digitale dell'europa*", 19 febbraio 2020.

96 Cfr. la scheda informativa presente sul sito della Commissione europea "*The Digital Services Act: ensuring a safe and accountable online environment",* in https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment_en. Per una ricostruzione delle proposte nella fase embrionale si rimanda a E. BROGI-I. NENANDIC, *European plan to increase transparency and accountability of the gatekeeper online platforms to protect democracy: EDMO`s role in the Commission's digital policy approach,* in *Medialaws.eu*, 18 dicembre 2020; cfr. A. PRETA, *DSA: obiettivi, criticità e nodi da sciogliere*, in *Medialaws.eu*, 7 dicembre 2020.

97 Cfr. O. POLLICINO, *Verso il Digital Services Act: problemi e prospettive. Presentazione del simposio*, in *Medialaws.eu*, 23 novembre 2020.

98 Art. 1 § 1 Regolamento (UE) 2022/2065 del Parlamento europeo e del Consiglio del 19 ottobre 2022 relativo a un mercato unico dei servizi digitali e che modifica la direttiva 2000/31/CE (regolamento sui servizi digitali) (da qui in poi indicato come "il Regolamento").

99 Cfr. *Considerando* n. 2 del Regolamento.

100 Cfr. R. BUCCA-M. SABATINI, *Digital Services Act, la Ue a una svolta: cosa cambia per utenti, aziende e big tech*, in *Agenda Digitale*, 19 maggio 2022.

101 V. G. DE GREGORIO, *L'alba di nuove responsabilità sulle piattaforme digitali: il Digital Services Act*, in *Agenda Digitale*, 15 dicembre 2020.

102 V. G. BAZZONI, *L'evoluzione normativa dell'intermediazione digitale: nuovi profili di responsabilizzazione*, in *Rivista italiana di informatica e diritto*, 1/2022, p. 201 ss.

gli articoli da 12 a 15 della direttiva *e-commerce*[103]. L'obiettivo primario è quello di "blindare", mediante la maggior cogenza derivante dall'atto di *hard law*[104], le responsabilità degli *information services* circa i fenomeni di violenza, *hate speech* e disinformazione online[105]. Questo ambizioso scopo viene sostanzialmente perseguito attraverso l'introduzione di una serie diversificata di obblighi e doveri di vigilanza per i diversi fornitori, modulata sulla base della tipologia del servizio di intermediazione offerto e della dimensione del soggetto operante (Capo II)[106]. In forza di tale quadro, i destinatari delle prescrizioni devono assicurare di aver posto in essere i rigidi adempimenti individuati dall'atto normativo affinché possano essere considerati esenti da responsabilità per i contenuti illegali che si trovino involontariamente ad ospitare[107]. Se, con riferimento ai più "rudimentali" servizi di *caching* e *conduit*[108], i vincoli possono definirsi sostanzialmente invariati, gli *hosting*, le *online platform* - quali i *social network* - e, in particolare, quelle che l'atto identifica come "*very large online platforms*"(VLP)[109] diventano destinatari di specifici adempimenti. In linea di principio, si può affermare che la regola chiave rimane quella di una generale irresponsabilità del fornitore fino al momento di venuta a conoscenza del contenuto illegale[110]: quest'ultimo fungerebbe da *dies a quo* dal quale emergerebbe un dovere di rapida attivazione per individuare la fonte, bloccare la disseminazione del contenuto nonché impedirne l'accessibilità. Il nuovo regime "a responsabilità condizionata" viene implementato mediante l'adozione di differenti misure. In primo luogo, vengono identificate procedure più celeri ed articolate per l'eliminazione di elementi o prodotti illegali, nonché per la presentazione di reclami e segnalazioni da parte degli utenti[111]. In secondo luogo, grava sulle tre

[103] Cfr. S. FLAMINIO, *Lotta alle fake news: dallo stato dell'arte a una prospettiva di regolamentazione per il "vivere digitale" a margine del Digital Service Act*, in *Rivista italiana di informatica e diritto*, 2/2022. È opportuno precisare che le novità descritte di seguito costituiscono proprio l'oggetto delle modifiche apportate agli articoli da 12 a 15.
[104] Cfr. F. MURONE, *Il Digital Service Act e il contrasto ai contenuti illeciti (pt. II)*, in *Iusinitinere.it*, 27 febbraio 2022.
[105] Cfr. G. RUOTOLO, *Le proposte europee di riforma della responsabilità dei fornitori di servizi su Internet*, in *Rivista italiana di informatica e diritto*, fasc. 1/2022, p. 19.
[106] Cfr. S. RUDOHRADSKÁ - D. TREŠČÁKOVÁ, *Proposals for the digital markets act and digital services act: broader considerations in context of online platforms*, in *EU and Comparative Law Issues and Challenges Series (ECLIC)*, 5/2021, p. 495 ss.
[107] Cfr. A. NICITA, *Le piattaforme online tra moderazione e autoregolazione: verso il Digital Services Act*, in *Medialaws.eu*, 25 novembre 2020.
[108] Cfr. art. 4 e 5 del Regolamento.
[109] In realtà è il soggetto pubblico a designare la piattaforma di notevoli dimensioni sulla base di un dato quantitativo. Difatti, in accordo all'articolo 33 § 4 la «Commissione, previa consultazione dello Stato membro di stabilimento o tenuto conto delle informazioni fornite dal coordinatore dei servizi digitali del luogo di stabilimento a norma dell'articolo 24, paragrafo 4, adotta una decisione che designa come piattaforma online di dimensioni molto grandi o motore di ricerca online di dimensioni molto grandi» quei soggetti con un numero di destinatari attivi pari o superiore a 45 milioni di utenti.
[110] V. G. ABALDO, *Una prospettiva di regolamentazione degli ISP attraverso il Digital Service Act*, in *Medialaws.eu*, 3 febbraio 2022.
[111] Art. 16 §1 stabilisce, infatti, che i «prestatori di servizi di memorizzazione di informazioni predispongono meccanismi per consentire a qualsiasi persona o ente di notificare loro la presenza nel loro servizio di informazioni specifiche che tale persona o ente ritiene costituiscano contenuti illegali. Tali meccanismi sono di facile accesso e uso e consentono la presentazione di segnalazioni esclusivamente per via elettronica». Sul punto cfr. M.R. ALLEGRI, *Il futuro digitale dell'Unione europea: nuove categorie di intermediari digitali, nuove forme di responsabilità*, in *Rivista italiana di informatica e diritto*, 2/2021, p. 14 ss.

categorie di servizio più sofisticate l'obbligo di informare l'autorità giudiziaria dello Stato membro nell'ipotesi di comminazione di un reato che possa comportare «una minaccia per la vita o la sicurezza di uno o più persone»[112]. All'interno di tale sistema prescrittivo, evidentemente congegnato per predisporre un celere e funzionale isolamento degli *harmful content*, acquista poi un significato essenziale il "segnalatore attendibile", una figura dotata di specifiche competenze nell'individuazione di contenuti dannosi o illegali, nonché di un'accertata indipendenza dalla piattaforma[113]. Bloccare il flusso di *illegal contents* in modo tempestivo e rigoroso risulta essere, dunque, uno dei principali doveri posti a carico dei prestatori di servizi più sofisticati, come si evince dalla disposizione che impone alle tre categorie già menzionate di *provider* di sospendere immediatamente «la prestazione dei loro servizi ai destinatari del servizio che con frequenza forniscono contenuti manifestamente illegali»[114].

Infine, in questo panorama di obblighi e di misure puntuali, la Sezione V introduce limitazioni ulteriori a carico delle piattaforme di notevoli dimensioni. Si prevede, ad esempio, che le *VLPs* – identificate sulla base di un criterio quantitativo – effettuino periodicamente una valutazione dei rischi sistemici connessi all'erogazione dei loro servizi, utilizzando come criterio la gravità e probabilità dell'evento. All'interno di tali ipotesi, di cui viene fornito un elenco apparentemente non tassativo, rientra, in prima battuta, proprio la «diffusione di contenuti illegali»[115]. Inoltre, è significativo segnalare come, in caso di inadempienza, gli Stati membri possano erogare delle sanzioni pecuniarie a carico delle grandi piattaforme che possono raggiungere anche il 6% del fatturato annuo[116].

Osservando l'intervento nella sua totalità, appare necessario affermare come il *Digital Service Act* costituisca un atto di considerevole impatto poiché realizza una riorganizzazione di portata epocale idonea a imprimere organicità ad una sfera che fino ad oggi pullulava di norme frammentate, regole asimmetriche nonché differentemente vincolanti. Rispetto alle criticità evidenziate nei paragrafi precedenti, dunque, la riforma potrebbe presentare rilevanti potenzialità.

Con riferimento al primario problema "sistemico" e di rilievo costituzionale prospettato, cioè il progressivo scivolamento di un potere decisorio notevole nelle mani di soggetti privati in materia di contrasto ai contenuti illegali *online*, si può dire che l'intervento si distingua per una compiuta enucleazione delle responsabilità dei fornitori. In particolare, l'azione dei colossi digitali appare riorganizzata all'interno di un nuovo e composito regime normativo che assoggetta l'azione dei prestatori di *intermediary services* ad

[112] Art. 18 §1.
[113] La figura del "*trusted flagger*", ai sensi dell'articolo 22 del Regolamento, viene individuata dal coordinatore dei servizi digitali dello stato membro e si qualifica per costituire un ente dotato di una particolare idoneità e competenza dell'identificazione di contenuti illegali.
[114] Art. 23 §1.
[115] Art. 34 §1 *lett.* a.
[116] Alle sanzioni è dedicato il Capo IV e, in particolare, l'articolo 52.

un solido controllo del soggetto pubblico. Ciò si evince in prima battuta dallo stesso art. 1 che identifica quale *main goal* del regolamento la creazione di un quadro armonizzato che definisca la responsabilità condizionata degli intermediari, chiamati ad operare secondo canoni di diligenza e di cooperazione con le autorità competenti[117].

In relazione ai più specifici nodi problematici evidenziati in merito al Codice anti *hate speech* del 2016, si può ritenere che l'eccessiva discrezionalità del prestatore nell'identificazione dell'illecito risulti parzialmente compensata dalla previsione posta dall'articolo 9. Quest'ultimo, infatti, specifica che «l'ordine di contrastare uno o più specifici contenuti illegali, emesso dalle autorità giudiziarie o amministrative nazionali competenti» debba contenere una serie tassativa di elementi, fra i quali primeggiano «la base giuridica» giustificante la misura, nonché una dettagliata motivazione a supporto contenente «un riferimento a una o più disposizioni specifiche del diritto dell'Unione o del diritto nazionale conforme al diritto dell'Unione»[118]. L'effettività di un controllo pubblico sull'operato degli intermediari appare poi rafforzata dalle numerose prescrizioni concernenti gli obblighi di trasparenza, a cui è dedicato il Capo III. Fra essi, risulta di particolare pregio la richiesta di pubblicazione di una dettagliata relazione annuale nella quale ogni *provider* dovrebbe rendere note tutte le informazioni concernenti l'attività di moderazione effettuata, una rendicontazione per cui la Commissione può individuare una formula standard in via esecutiva[119]. In aggiunta a tali adempimenti richiesti dall'articolo 15, l'articolo 24 esige poi che le piattaforme comunichino le segnalazioni pervenute aventi ad oggetto la presenza di contenuti illegali, specificando altresì l'avvenuto accertamento di notifiche manifestamente infondate nonché le eventuali controversie sottoposte a meccanismi di risoluzione extragiudiziale. Queste informazioni devono essere comunicate al *Digital Service coordinator* o alla Commissione stessa, i quali hanno facoltà di esigere un'ulteriore integrazione delle informazioni[120]. Inoltre, con riferimento alle *VLPs*, i soggetti pubblici possono avanzare delle richieste di accesso ai dati con il fine di sindacare la conformità del servizio erogato alle norme previste dal regolamento. Infine, date le rilevanti implicazioni connesse all'uso dello strumento algoritmico segnalate nel paragrafo precedente, è interessante porre in luce come le rigorose previsioni concernenti il monitoraggio da parte dell'autorità si estendano altresì anche al

[117] Si veda il *Considerando* n. 3 il quale afferma «un comportamento responsabile e diligente da parte dei prestatori di servizi intermediari è essenziale per un ambiente online sicuro, prevedibile e affidabile e per consentire ai cittadini dell'Unione e ad altre persone di esercitare i loro diritti fondamentali garantiti dalla Carta dei diritti fondamentali dell'Unione europea («Carta»), in particolare la libertà di espressione e di informazione, la libertà di impresa, il diritto alla non discriminazione e il conseguimento di un livello elevato di protezione dei consumatori».

[118] Art. 9 § 2. Ulteriore precisione è richiesta altresì dall'articolo 10 concernente la richiesta di ulteriori informazioni da parte dell'autorità pubblica.

[119] Art. 15.

[120] Art. 24 §3.

sistema di IA utilizzato dalla piattaforma e, nello specifico, possono riguardare la «progettazione», la «logica», il «funzionamento» e la «sperimentazione» dell'algoritmo adoperato[121].

L'insieme di misure sinora menzionate, contribuendo a sottoporre l'azione degli intermediari ad un penetrante controllo e al rispetto di solidi requisiti di trasparenza, possono dunque contenere al suo interno dei preziosi correttivi all'impostazione contenuta nell'atto di *soft law,* una ristrutturazione organica idonea ad attenuare i più vistosi disequilibri concernenti l'azione del soggetto privato nella lotta per la rimozione dei contenuti illegali e, dunque, anche nella diffusione di contenuti qualificabili come *hate speech*. Effetti virtuosi scaturenti dall'entrata in vigore del regolamento potrebbero altresì emergere se si analizza il tema attraverso l'altra prospettiva di risonanza costituzionale tracciata, quella proiettata verso la valutazione dell'impatto di una non opportunatamente regolata *content moderation* a carico dei diritti fondamentali dell'utente e, più specificamente, per la libertà di espressione dello stesso. In primo luogo, è innegabile che la centralità che il criterio della trasparenza assume all'interno del provvedimento acquisisca una pervasività tale da esercitare una diretta ricaduta sulle posizioni soggettive dei destinatari[122]. In primo luogo, colpisce – proprio in forza della sua valenza orizzontale priva di eccezioni - l'obbligo, gravante su ogni categoria di prestatori, di rendere chiare, inserendole all'interno delle proprie condizioni generali, «le politiche, le procedure, le misure e gli strumenti utilizzati ai fini della moderazione dei contenuti, compresi il processo decisionale algoritmico e la verifica umana, nonché le regole procedurali del loro sistema interno di gestione dei reclami»[123]. Tale insieme di regole deve risultare accessibile attraverso «un linguaggio chiaro, semplice, comprensibile, facilmente fruibile e privo di ambiguità». L'obiettivo di fissare, in merito all' attività di moderazione concernente contenuti illegali, regole intellegibili e solidamente ancorate al rispetto del principio di legalità sembra inoltre essere assolto dalla "proceduralizzazione" e articolazione puntuale dei meccanismi di segnalazione e reclamo posti dall'articolo 16, nonché dalle garanzie che lo accompagnano. Fra di esse colpisce la previsione che esige che la decisione di sospensione o cessazione (temporanea o parziale) del servizio, eventualmente disposta in seguito all'accertamento di una attività illecita, sia accompagnata da una adeguata motivazione. Quest'ultima, fra i vari elementi, dovrebbe contenere proprio una puntuale indicazione del fondamento giuridico nazionale o euro-unitario utilizzato dal prestatore[124].

Nell'ottica di promuovere un'attenuazione della discrezionalità dei fornitori in favore dei diritti dell'utente, appaiono poi di assoluto rilievo le garanzie poste dagli articoli 20, 21 e 23 del regolamento.

[121] Art. 40 §3.
[122] Cfr. G. FROSIO, *Platform Responsibility in the Digital Services Act: Constitutionalising, Regulating and Governing Private Ordering*, 3 ottobre 2022, Forthcoming in A. SAVIN- J. TRZASKOWSKI (a cura di), *Research Handbook on EU Internet Law*, Edward Elgar, 2023, available at SSRN: https://ssrn.com/abstract=4236510, pp. 14-15.
[123] Art. 14 §1.
[124] Art. 17 § 3 *lett.* d.

L'articolo 20, infatti, introduce l'obbligo, per le piattaforme online, di procedere all'istituzione di un sistema interno di gestione dei reclami presentati dai destinatari del servizio in merito a segnalazioni precedenti, o contro le decisioni della piattaforma di sospensione, disabilitazione o definitiva cessazione dell'account per un periodo di almeno sei mesi. La previsione impone che il fornitore tratti l'eventuale reclamo «in modo tempestivo, non discriminatorio, diligente e non arbitrario» qualora emergano motivazioni sufficientemente attendibili che supportino il ricorso dell'utente[125]. Inoltre, l'eventuale sospensione del servizio in presenza di abusi, disciplinata dall'articolo 23, richiede comunque una valutazione caso per caso che vada a contemplare anche la sistematicità e reiterazione della violazione: entrerebbero in gioco, dunque, i già evocati parametri di obiettività, tempestività, diligenza e proporzionalità, uniti a un giudizio fondato su un criterio numerico e sulla gravità del contenuto prodotto. Nella medesima prospettiva, appare ugualmente essenziale, infine, la norma che introduce un diritto di accesso a meccanismi di risoluzione extragiudiziale delle controversie di fronte ad organismi la cui idoneità è accertata dal coordinatore dei servizi digitali in cui è stabilito l'organo giudicante[126].

Alla luce delle considerazioni tracciate, è sicuramente da accogliere con favore lo sforzo del legislatore euro-unitario di realizzare un delicato contemperamento fra l'obiettivo primario posto alla base dell'intervento - il consolidamento di una dimensione virtuale immune alla proliferazione di contenuti illeciti[127]-, e l'esigenza, dotata del medesimo peso assiologico in un sistema che aspira comunque a conformarsi sulle basi fondanti della *rule of law*[128], che tale opera di innalzamento delle garanzie non si traduca in un *vulnus* collaterale a danno dei diritti fondamentali dell'utente. Tale aspirazione non si coglie soltanto dai ricorrenti ed espliciti riferimenti a valori nodali, dal principio della certezza del diritto a quello di uguaglianza, ma si deduce implicitamente dalle dettagliate garanzie procedurali, disseminate nell'atto, che ricalcano da vicino i fondamenti tipici dello Stato di diritto costituzionale, dall'obbligo di motivazione della decisione al diritto di appello di fronte ad un soggetto indipendente e imparziale. Tali principi concretano, seppur con manifeste differenze giustificate dalla peculiarità del contesto, un diritto ad un rimedio "giurisdizionale" accessibile ed effettivo.

Tuttavia, nonostante tali elementi positivi evidenziati, appare comunque opportuno segnalare come le essenziali criticità e i rischi evidenziati in merito all'applicazione del codice di condotta rischino comunque

[125] Art. 20 § 4. La previsione specifica, inoltre, che «se un reclamo contiene motivi sufficienti per indurre il fornitore della piattaforma online a ritenere che la sua decisione di non dare seguito alla segnalazione sia infondata o che le informazioni oggetto del reclamo non siano illegali né incompatibili con le condizioni generali, o se tale reclamo contiene informazioni indicanti che il comportamento del reclamante non giustifica le misure adottate, il fornitore della piattaforma online annulla senza indebito ritardo la decisione di cui al paragrafo 1».

[126] Art. 21 §3.

[127] Cfr. F. MURONE, *Il Digital Service Act e il contrasto ai contenuti illeciti online*, in *Iusinitinere.it*, 21 dicembre 2021.

[128] Cfr. G. DE GREGORIO, *The Digital Services Act: A Paradigmatic Example of European Digital Constitutionalism*, in *Diritti Comparati*, 17 maggio 2021.

di permanere nella nuova realtà normativa rimodellata dal *Digital Service Act*. Invero, il primo elemento critico evidenziato in relazione all'atto di *soft law*, quello della problematicità definitoria, non sembra essere superato dal nuovo *framework* regolatorio. A ben vedere, infatti, il regolamento non compie alcuno sforzo chiarificatore significativo nel dettare quali siano le tipologie di contenuti qualificabili come "illegali" - fra cui rientrano, ovviamente, anche i discorsi d'odio – ma si limita, semplicemente, a sugellare l'importanza del fondamento legale nei vari interventi decisori, il quale non può che adagiarsi, ancora una volta, sulle fonti euro-unitarie e nazionali pertinenti[129]. In tal modo, quindi, andrebbe a riproporsi quel meccanismo distorsivo "di rimandi" normativi, già analizzato, inidoneo ad essere risolutivo in merito alla fumosità della fattispecie[130]. In ultima battuta, quindi, l'impostazione di numerose previsioni appare ancora una volta vocata a ribadire l'imprescindibilità di una solida "*legal basis*" nella procedura di adozione di una decisione autoritativa, senza tuttavia spingersi fino a riempire il concetto di illecito un contenuto sostanziale autonomo.

Analoghe perplessità emergono in merito ai limiti e alle criticità connaturate all'uso dello strumento algoritmico per i diritti fondamentali dell'utente e, in particolare, per alcune categorie minoritarie[131]. Risulta indubbiamente percepibile l'accresciuta sensibilità del legislatore europeo in merito a tali aspetti, come testimonierebbe il riferimento ricorrente al principio di non discriminazione, un'aggiunta particolarmente valorizzata in fase di emendamento, sia i numerosi riferimenti alla necessità di sottoporre anche i sistemi di intelligenza artificiale utilizzati per l'attività di moderazione ad una logica di trasparenza e conoscibilità[132]. Tuttavia, lo sforzo di incasellare anche la parte più tecnica e impenetrabile della *content moderation* in una cornice di legalità non appare di per sé sufficiente ad innescare un radicale ripensamento nell'uso dei sistemi automatizzati realmente improntato ad una logica di tutela sostanziale. La previsione, a carico delle piattaforme, di procedere alla sperimentazione e all'adeguamento dei loro sistemi algoritmici come misura di attenuazione dei rischi sistemici connessi all'uso della piattaforma non appare, infatti, risolutiva [133]. Tale indicazione sembra arrestarsi in superficie senza scalfire i principali nodi connaturati alla moderazione automatizzata.

In conclusione, ciò che affiora in relazione ai due fattori problematici esaminati sembra confermare il fatto che l'intervento titanico perseguito dal legislatore continentale di "porre le briglie" al dilagante potere delle piattaforme, mediante un'onnicomprensiva articolazione delle garanzie, rimanga esso stesso

[129] M. HUSOVEC- I. ROCHE LAGUNA, *Digital Services Act: A Short Primer*, July 5, 2022 in M. HUSOVEC- I. ROCHE LAGUNA (eds.), *Principles of the Digital Services Act*, Oxford University Press, Forthcoming 2023, disponibile a https://ssrn.com/abstract=4153796 or http://dx.doi.org/10.2139/ssrn.4153796.
[130] Cfr. G. DE MINICO, *Fundamental rights, european digital regulation and algorithmic challenge*, in *Medialaws.ue*, 1/2021, p. 21.
[131] Cfr. M.C. FALCHI, *Intelligenza Artificiale: se l'algoritmo è discriminatorio*, in *Iusintinere.it*, 5 ottobre 2020.
[132] V. E. GARZONIO, *L'algoritmo trasparente: obiettivi ed implicazioni della riforma dello Spazio digitale europeo*, in *Rivista italiana di informatica e diritto*, 2/2021, p. 28 ss.
[133] Art. 35.

imbrigliato in una logica di espansione delle tutele di natura meramente procedurale[134]. In un contesto globale in cui il costituzionalismo digitale *in fieri* si confronta ancora con divergenze sensibili intorno ai principi fondanti, tale approccio rappresenta certamente la via più percorribile per costruire una soddisfacente sistema di salvaguardia dei diritti fondamentali nel cyberspazio[135]. Tale compromesso, tuttavia, se non accompagnato da una robusta azione sul fronte sostanziale, rischia di creare dei presidi garantistici formalmente imponenti ma vuoti[136], perché ancora esposti alla discrezionalità interpretativa e applicativa delle piattaforme, libere, ancora una volta, di plasmarne il significato[137].

4. Osservazioni conclusive

La strategia di contrasto all'*hate speech* nella dimensione virtuale non può che essere osservata come uno dei "precipitati concreti" più tangibili del processo di consolidamento del cosiddetto costituzionalismo digitale all'interno della sfera europea. Tale fenomeno ha segnato il progressivo abbandono dell'approccio liberista, condiviso con il legislatore d'oltreoceano[138], in favore di un nuovo assetto normativo in cui il paradigma economicistico posto alle radici dell'ordinamento potesse coniugarsi con la tutela dei diritti fondamentali[139]. Questa riconfigurazione non può essere ridotta, tuttavia, soltanto ad un percorso di progressiva apertura verso istanze personaliste in tale ambito. Invero, in essa può intravedersi una visione consapevole, da parte delle Istituzioni europee, in merito al fatto che l'affermazione dei poteri privati si stesse sviluppando a detrimento delle basi democratiche del sistema e in violazione dei principi dello Stato di diritto[140]. Per tal ragione, tale cambio di rotta, caratterizzato da svolte normative chiave e dal

134 M. BETZU, *Poteri pubblici e poteri privati nel mondo digitale*, in *Rivista "Gruppo di Pisa"*, fasc. 2/2021, pp.180-181.
135 O. POLLICINO, *Libertà di espressione, piattaforme digitali e cortocircuiti di natura costituzionale*, in *Privacy&*, 1/2021, p. 7.
136 G. PALOMBELLA, *È possibile una legalità globale?,* Bologna, Il Mulino, 2012, nel descrivere le modalità attraverso le quali il principio della *rule of law* si modella a livello globale descrive, appunto, un'espansione del concetto di *accountability*, poiché, «in mancanza dei canali disponibili nelle democrazie costituzionali» diventano centrali «gli obblighi di adeguata motivazione, trasparenza, responsabilità» (p. 161).
137 V. C. CAUFFMAN- C. GOANTA, *A New Order: The Digital Services Act and Consumer Protection, in European Journal of Risk Regulation,* vol. 12, issue 4, 2021, p. 758 ss.
138 Si allude a quell'approccio regolatorio accolto nella prima fase in cui le istituzioni europee hanno dovuto confrontarsi con l'emersione degli intermediari online e delle potenzialità economiche del settore digitale. Questo filone, noto come "*digital liberalism*", trova nell'impostazione della cosiddetta direttiva *e-commerce* la sua principale traduzione normativa e nel modello americano, confluito nel noto *Communication Decency Act* e dal *Digital Millennium Copyright Act*, il suo principale presupposto ispiratore. La strategia liberista, volta ad incentivare uno sviluppo del settore privo di ostacoli, è ben rappresentata dalla nota *Section 230* del *Decency Act* e ha costituito un approccio chiave nel processo di affermazione dei grandi colossi digitali, dimostrando ancora una volta la sempiterna centralità del primo emendamento nel costituzionalismo americano. Sul punto si vedano le riflessioni di G. DE GREGORIO, *Digital constitutionalism in Europe*, Cambridge, Cambridge University press, 2022, p. 41 ss.
139 Cfr. E. CELESTE, *Digital constitutionalism: a new systematic theorisation*, in *International Review of Law, Computers & Technology*, 2019, p. 81.
140 Cfr. M. MOORE- D. TAMBINI (a cura di), *Digital Dominance: The Power of Google, Amazon, Facebook, and Apple*, Oxford, Oxford University Press, 2018. D'altra parte, tale presa di coscienza, all'interno del contesto statunitense, non è riuscita comunque a scalfire del tutto quelle diffuse resistenze all'introduzione di una regolamentazione dell'operato delle piattaforme a difesa della libertà di espressione dei singoli. Dietro tale riluttanza non emergerebbe soltanto il già

ruolo pioneristico svolto dalla Corte di Giustizia, si può considerare *in primis* come un più ampio processo che muove verso l'inquadramento della sfuggente azione delle *Big Tech* dell'informazione digitale entro la cornice della *rule of law*[141]. Alla luce di tali presupposti, il rapporto fra soggetto privato e soggetto pubblico risulta profondamente mutato nello scenario attuale[142]. I prestatori di servizi non solo divengono destinatari passivi di regole uniformi attraverso percorsi regolatori che inizialmente si fondano su base volontaristica e sul *soft law*; quanto, piuttosto, assumono anche le vesti di *private enforcers* della legalità costituzionale, sviluppando una proficua sinergia con il decisore pubblico[143]. Tale cooperazione è divenuta centrale anche nel contrasto ai contenuti illegali e, più nello specifico, nella lotta alla diffusione di contenuti d'odio online, come dimostra il *Code of Conduct on Combatting Illegal Hate Speech Online*, utilizzato come caso studio all'interno di tale riflessione. Tuttavia, com'è stato sottolineato, nelle pieghe di tale modello si celano intrinseche problematiche che si manifestano negli aspetti applicativi. Il ruolo attuativo svolto dalle grandi piattaforme risulta infatti tollerabile solo nella misura in cui esse agiscano come meri esecutori di una decisione che giace nelle mani del decisore statuale, e in assenza di significativi margini discrezionali, soprattutto in un ambito particolarmente delicato come quello afferente all'*hate speech*. Il descritto equilibrio sembra mancare nella configurazione dell'atto normativo oggetto di analisi in quanto alcuni suoi elementi costitutivi affiderebbero all'ente privato il compito di compiere gli opportuni bilanciamenti fra i valori in gioco, in assenza di un adeguato controllo pubblico e democratico. Più nel

menzionato approccio liberale – e liberista – che vede nell'eccesso di regolazione un ostacolo all'esercizio di un'attività economica, quanto una determinata visione della libertà di espressione e della sua vincolatività in senso orizzontale. Per anni, dunque, sarebbe prevalsa una contrarietà, in dottrina e nella giurisprudenza della Corte Suprema, all'applicazione del limite della libertà di manifestazione del pensiero nei confronti dei soggetti privati se non nelle ipotesi limitate in cui quest'ultimi vadano a configurarsi come "*state actors*" e, quindi, come soggetti idonei a esercitare un controllo penetrante ed effettivo sul discorso pubblico, concretizzabile come un vero e proprio potere censorio che impedisse al cittadino di individuare, ragionevolmente, canali di espressione alternativi. Questa interpretazione ed applicazione del I emendamento, riscontrabile in relazione al caso delle cosiddette *company towns* (cfr. il noto caso *Marsh v. Alabama*, 1946), costituisce, comunque, un approccio eccezionale. Nonostante permanga un consenso maggioritario verso tale visione, di recente si assiste ad un ampio dibattito sulla tenuta di tale sistematizzazione in presenza di colossi del digitale che tendono ad assumere un ruolo sempre più simile al potere esercitato da un soggetto pubblico e a qualificarsi come "*modern public squares*" (*Packingham v. North Carolina*, 2017). All'interno di tale scenario, si inizia ad affermare l'idea che nelle piattaforme si stia condensando una considerevole quantità di informazione, con conseguenti capacità di controllare il discorso pubblico, riflessioni emerse in particolare nelle *concurring opinion* del giudice Thomas in *Packingham* e, sulla scia di quest'ultima, nella più recente sentenza *Joseph R. Biden, Jr., President of the United States, et al. v. Knight First Amendment Institute at Columbia University, et al. 593 U. S.* del 2021. Per una ricostruzione ampia si rimanda a M. MONTI, *La Corte Suprema statunitense e il potere delle piattaforme digitali: considerazioni sulla privatizzazione della censura a partire da una concurring opinion*, in *DPCE online*, 1/2021, p. 2781 ss.; cfr., altresì, R. NIRO, *Piattaforme digitali e libertà di espressione fra autoregolamentazione e coregolamentazione: note ricostruttive,* in *Osservatorio sulle fonti*, 3/2021, p. 1369 ss.

[141] G. DE GREGORIO, *Digital constitutionalism across the Atlantic*, in *Global constitutionalism*, 11/2022, p. 297 ss. Per un approfondimento inerente al ruolo cruciale assunto della Corte di Giustizia in tale percorso cfr. G. DE BURCA, *After the EU Charter of Fundamental Rights: The Court of Justice as a Human Rights Adjudicator?*, in *Maastricht Journal of European and Comparative Law*, 2013.

[142] J. BALKIN, *Old-school/new-school speech regulation*, cit, pp. 2298-2299.

[143] Originariamente tale "alleanza" era inquadrata con l'eloquente metafora dell' «*inivisible handshake*», in M.D. BIRNHACK-N.ELKIN-KOREN, *The Invisible Handshake: The Reemergence of the State in the Digital Environment*, in *Virginia Journal of Law and Technology*, 2003, p. 14 ss.

dettaglio, la vaghezza definitoria, da un lato, e lo strumento algoritmico utilizzato, dall'altro, contribuirebbero ad incentivare l'attività censoria delle piattaforme con conseguenze percettibilmente negative per la libertà di espressione degli utenti. Valutato a sei anni di distanza dalla sua sottoscrizione, dunque, il *Code of Conduct on Combatting Illegal Hate Speech Online* non sembra porsi pienamente in linea con i presupposti chiave del costituzionalismo digitale. Se, infatti, quest'ultimo si distingue per un'assimilazione dello "sregolato" potere delle *online platforms* in una cornice pubblicistica imperniata sulla *governance* democratica e sulle garanzie individuali, lo scivolamento nelle mani di quest'ultime di una considerevole discrezionalità decisoria sembra riprodurre un disequilibrio da cui, ancora una volta, è il principio della *rule of law* ad uscirne sofferente.

Le problematicità connaturate al Codice di condotta sono state indagate alla luce del nuovo *framework* normativo caratterizzato dalla definitiva approvazione del *Digital Services Act* ai fini di comprendere se la disciplina organica introdotta da quest'ultimo potesse costituire un superamento delle problematiche emergenti nella disciplina di *soft law*, generando così una strategia di contrasto ai discorsi d'odio che fosse efficace e, al contempo, capace di assoggettare l'azione esecutiva delle piattaforme ai principi dello Stato di diritto. In linea generale, è stato osservato come l'intervento tenti meritoriamente di dar concretezza all'imperativo della Commissione proiettato verso la creazione di uno spazio virtuale sicuro in cui "ciò che è vietato offline sia vietato online"[144]. Tale aspirazione viene implementata mediante un provvedimento complesso ed articolato nel quale la responsabilità e l'attività dei prestatori viene sottoposta a ferrei obblighi di diligenza e trasparenza. Quest'ultima, in particolare, si erge a limite che vincola l'operato della piattaforma nella sua totalità – soprattutto a quelle di notevoli dimensioni – estendendosi anche al funzionamento dello strumento algoritmico. Contestualmente, la tutela dei diritti fondamentali diviene centrale nell'impostazione del provvedimento, concretizzandosi in una serie di garanzie procedurali in favore dell'utente. Emergono, dunque, una pluralità di elementi di significativa rilevanza nell'innovativo atto di *hard law*, i quali danno contezza di una volontà di responsabilizzazione delle piattaforme e di cooperazione con il soggetto privato nella eliminazione dei contenuti illegali, nell'ottica della legalità e di un corretto equilibrio fra quei valori in gioco che vengono in rilievo in un settore in cui posizioni soggettive e interessi economici crescenti si intrecciano. Nonostante ciò, tali fattori sembrano inidonei, di per sé, a superare le principali criticità connaturate al Codice di condotta. Dalle disposizioni del Regolamento non sembra emergere né una definizione autonoma e articolata di "contenuto illegale" tale da circoscrivere e indirizzare l'azione delle piattaforme in tale ambito, né una

[144] Questo costituiva l'obiettivo primario del *Digital Service Package*, cfr. https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package. Si veda, in particolare, il testo della Comunicazione "*Plasmare il futuro digitale dell'Unione*, COM(2020) 67 final.

ripianificazione dell'utilizzo dello strumento algoritmico tale da poter rimediare con successo alle principali falle di funzionamento e alle loro ricadute sulla libertà di espressione e sui gruppi minoritari.

Guardando, dunque, all'intervento nella sua totalità e complessità, quello che sembra profilarsi, nonostante due anni di lunghe trattative volte ad affinare lo strumento di *hard law* ai fini di rendere tale sistema sensibile alle istanze di tutela sostanziale, è un prodotto normativo non in grado di fugare del tutto i rischi di censura collaterale già segnalati. Se si presta attenzione allo spirito sotteso alla disciplina, non è da escludere che l'effetto censorio possa addirittura accentuarsi. Invero, benché il regolamento si distingua per essere una normativa ispirata alla volontà di inquadrare l'azione degli intermediari all'insegna dei principi fondanti del diritto europeo e della Carta di Nizza, - e, quindi, ad un agire «diligente, obiettivo, non discriminatorio e proporzionato, tenendo debitamente conto dei diritti e degli interessi legittimi di tutte le parti coinvolte e fornendo le necessarie garanzie contro la rimozione ingiustificata di contenuti legali»[145]- la lotta ai contenuti illeciti sembra comunque rimanere il principale obiettivo posto a fondamento dall'intervento e la primigenia vocazione dell'atto. Più chiaramente, tale finalità, dunque, risulterebbe in prima battuta la modalità per fornire effettiva concretizzazione alla tutela delle posizioni soggettive coinvolte nello spazio digitale. Proprio intorno ad essa tenderebbe a modellarsi il complesso regime di responsabilità che, come più volte sottolineato, diventa particolarmente stringente per le *very large platforms*. Il complesso di vincoli procedurali che condizionano l'operato degli intermediari, dagli obblighi di trasparenza alle periodiche valutazioni dei rischi sistemici, appare comunque funzionale al raggiungimento del primario scopo dell'intervento, quello di rafforzare il contrasto all'illecito. Analogamente, le garanzie procedurali istituite a tutela dell'utente risultano logicamente secondarie rispetto allo scopo primario che muove il legislatore europeo. Per tali ragioni, una conseguenza non voluta - ma probabile - che potrebbe verificarsi è che la maggior cogenza derivante dal tipo di fonte utilizzata e il rigido sistema di obblighi stringenti introdotto dal regolamento possano in realtà incentivare una politica volta all'eliminazione indiscriminata di elementi qualificabili come "illegali"[146]. Più nel dettaglio, anche all'interno del nuovo quadro regolatorio andrebbero a riproporsi le criticità sollevate in merito al funzionamento del meccanismo del *notice and takedown*: in presenza di doveri e sanzioni, lo sforzo di svolgere un'attività di moderazione scevra da errori di valutazione viene sovrastato dall'esigenza di porre in essere un blocco sistematico dei contenuti sospetti[147]. L'articolato sistema di controllo posto a carico delle *Internet platforms* potrebbe dunque incentivare quel paradigma operativo concretizzabile nell'idea del "*shoot first, ask questions later*"[148] che, seppur accompagnato un inedito sistema di garanzie, rimarrebbe

[145] *Considerando* n. 26 del Regolamento.
[146] Cfr. J. BALKIN, *Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation*, in *UC Davis Law Review*, 2018, pp. 1176-1177.
[147] Cfr. S.K. KATYAL, *The New Surveillance*, in *Case Western Reserve Law Review*, 2003, p. 297 ss.
[148] S. F. KREIMER, *Censorship by Proxy*, cit., p. 28, nota 52.

comunque dominante. In conclusione, pur dovendo inevitabilmente attendere i primi riscontri applicativi, la logica normativa prevista all'interno del DSA non consente di giudicare la *collateral censorship*, anche in uno scenario normativo sensibilmente rinnovato, un fenomeno problematico definitivamente superato[149].

[149] Cfr. P. DUNN, *Il contrasto europeo all'hate speech online: quali prospettive future?*, in *MediaLaws.eu*, 20 gennaio 2021.